University of Alberta

METHODS FOR MIXED OUTCOME DATA

by

ALEXANDER R. DE LEON

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

in

Statistics

Department of Mathematical and Statistical Sciences

Edmonton, Alberta
Fall 2002

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-81177-8

Canada

University of Alberta

Library Release Form

Name of Author: Alexander R. de Leon

Title of Thesis: **METHODS FOR MIXED OUTCOME DATA**

Degree: Doctor of Philosophy

Year this Degree Granted: 2002

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Alexander R. de Leon

Alexander R. de Leon
CAB 632
Department of Mathematical & Statistical Sciences
University of Alberta
Edmonton, AB
Canada T6G 2G1

Date: Sept. 3, 2002

. . . Extend yourself —
it is the Nile, the sun is shining,
everywhere you turn is luck.

— Louise Gluck
(from *The Undertaking*)

UNIVERSITY OF ALBERTA

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Methods for Mixed Outcome Data** submitted by **Alexander R. de Leon** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in Statistics.

Dr. Narasimha Prasad (Chair)

Dr. K. C. Carrière (Supervisor)

Dr. Rohana Karunamuni

Dr. Stephen Newman (Public Health Sciences)

Dr. Edward Bedrick (University of New Mexico)

July 5, 2002

To my family

Abstract

The focus of this thesis is on developing new methods for the analysis of multivariate data with mixtures of multinomial, ordinal, and continuous variables. A model-based approach is taken that relies on specifying a model for the joint distribution of the variables. This approach provides a systematic and non-*ad hoc* way of analyzing mixed data, as opposed to those currently available.

By adopting a general location model (Olkin and Tate, 1961) for mixed data, exact global likelihood ratio tests of the so-called location hypotheses are obtained, both in one-sample and multi-sample settings. For the one-sample case, it is shown that the likelihood ratio test is consistent and unbiased. In addition, simulation studies show that it performs quite competitively relative to the approach that carries out separate but simultaneous tests of the location parameters. Extensions to the multi-sample case are also studied, and an attempt is made to incorporate heterogeneity in the tests.

As another alternative to maximum likelihood estimation, the pairwise likelihood approach is adapted to the grouped continuous and conditional grouped continuous models. Unlike maximum likelihood estimation, the proposed method is computationally simple, and unlike partition maximum likelihood methods (Bedrick *et al.*, 2000; Poon and Lee, 1987), there is no need

to deal with multiple sets of estimates. The estimators based on the proposed method are shown to be consistent and asymptotically normally distributed, and tend to have minimal bias and mean-squared errors.

A general model for mixed multinomial, ordinal and continuous data, called the *general mixed-data model*, is also developed. In contrast to existing models for mixed data, the proposed model not only accounts for the different measurement levels in the data but also incorporates associations between the three variable types. Maximum likelihood and maximum pairwise likelihood methods are outlined for the model, with the latter providing a more computationally feasible alternative to the former. The asymptotic distributions of the corresponding estimators are also derived.

Finally, a generalized Mahalanobis distance for mixed data from several populations is proposed, by applying the Kullback-Leibler divergence to the general mixed-data model. Asymptotic distributions of the distance under the hypothesis of non-distinct groups are derived, and large-sample tests of hypotheses are constructed. Simulation studies suggest the tests to be well-behaved in finite samples.

Acknowledgements

Prof. K. C. Carrière has been a constant source of energy and inspiration for me. I thank her for nudging me over the edge, for her concern and patience that go over and beyond the call of duty. Special thanks go to Profs. E. J. Bedrick, R. J. Karunamuni, and S. C. Newman, for kindly agreeing to serve on my examination committee, and to Prof. N. G. N. Prasad, for chairing it. Thanks as well to Prof. P. M. Hooper, for chairing my candidacy committee. My *utang na loob* is especially great to Profs. R. J. Karunamuni, N. G. N. Prasad, D. P. Wiens, and Y. Wu, who taught me statistics and helped me in more ways than they can imagine.

My thanks and gratitude to the Department of Mathematical and Statistical Sciences, for providing me with financial assistance at various stages of my studies, and the Alberta Heritage Foundation for Medical Research, for awarding me a generous studentship that enabled me to work full-time on my research.

Due thanks to the following people who fed me, bought me coffee, told me a lot of stories, statistical and otherwise, lent me their books, edited me, corrected me—and thus, taught me—throughout my years in Edmonton:

Febe and Sam, for their words and laughter of pure joy, and for always tolerating my excesses with good humour;

Caetano, for his kindness and friendship;

Manuela, for the innumerable Portuguese dinners of *bacalhau* and *batata frita* I enjoyed at her place, for teaching me *sueca*, but most of all, for her caring and generosity;

Abdul—who knows everything but does not blink—for all his help, technical and otherwise;

Ella, of course;

My university friends and classmates: Eshetu A., Eshetu W., Eunha, Giseon, Julie, Melody, Rong, Victor, William and Christine, and Xiaoming, for the company;

My *Pinoy barkada*: Allan, Alex, Chito, and Jesser—life being what it is these days, I think it'd be wiser to be discreet;

Ate Thess, Nonie, Mommy Celi, Simone, and Vince, with whom I spent two memorable summers spreading mayhem in Japan;

My aunts, uncles and cousins in Canada and the U. S., for always taking care of me during the few times I visited Toronto and Florida;

And my family: Mama and Papa, Kiyad,Ate Nene, Kuya Eric and Leng, Kuya Freddie and Jindra, Tetet, and my nieces and nephews, who show me that even if we are all becoming weirder as the years pass, there will always be home.

Alex de Leon
Edmonton, AB
July 2002

Contents

1 Introduction **1**

1.1 Background of the Thesis 1

1.2 Issues in the Analysis of Mixed Data 3

1.2.1 Global Testing of Mixed Data Hypotheses 4

1.2.2 Computationally Efficient Estimation Methods 8

1.2.3 Models for Mixed Nominal, Ordinal and Continuous Data 11

1.2.4 Multivariate Methods for Mixed Data 13

1.3 Overview of the Thesis 16

2 Location Hypothesis Tests for Mixed Data **18**

2.1 Introduction 18

2.2 General Location Model 21

2.2.1 Likelihood Function 23

2.2.2 Maximum Likelihood Estimation 24

2.3 One-Sample Location Hypotheses in the Mixed Data Case . . 27

2.3.1 Case of Known Covariance Matrix 28

2.3.2 Case of Unknown Covariance Matrix 32

2.3.3 Properties of the Likelihood Ratio Test 37

2.3.4 Simulation Results 38

2.4 Extension to G-Sample Case 41

2.4.1 Case of Complete Homogeneity 43

2.4.2 Heterogeneous Cases 55

2.5 Discussion 58

3 Pairwise Likelihood Approach to Grouped Continuous Model and Its Extension **65**

3.1 Introduction 65

3.2 Grouped Continuous Model 69
3.3 Maximum Likelihood Estimation 70
3.4 Maximum Pairwise Likelihood Estimation 72
3.5 Asymptotic Results 75
3.6 Simulation Study 78
3.7 Conditional Grouped Continuous Model: Extension to Mixed Ordinal and Continuous Data 81
3.8 Discussion 88

4 General Mixed-Data Model: Extension of General Location and Grouped Continuous Models 94
4.1 Introduction 94
4.2 General Mixed-Data Model 99
4.2.1 Case with $C = L = Q = 1$ and $S = 2$: An Example . . 105
4.3 Maximum Likelihood Estimation 107
4.3.1 Case of a Single Ordinal Variable ($Q = 1$) 108
4.3.2 General Case ($Q \geq 2$) 111
4.4 Maximum Pairwise Likelihood Estimation 114
4.5 Asymptotic Distributions of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^{PL}$ 117
4.6 Statistical Inference 120
4.7 Appendicitis Data Example 122
4.8 Discussion 125

5 A Generalization of Mahalanobis Distance to Mixed Qualitative and Quantitative Data 130
5.1 Introduction 130
5.2 A Generalized Mahalanobis Distance 132
5.3 Asymptotic Results 137
5.4 Simulation Study 140
5.5 Example 142
5.6 Discussion 143

6 Concluding Remarks 148
6.1 Summary 148
6.2 Future Research 153

Bibliography 156

A Review of Optimization Methods 171

A.1 Determining the Direction 171

A.2 Choosing a Step Size 172

A.3 Survey of Basic Methods 172

B S-PLUS Programs Used in the Thesis 174

B.1 Calculation of Critical Values in Table 2.1 174

B.2 Calculation of Power Values in Tables 2.2 and 2.3 175

B.3 Calculation of Bias and RMSE in Tables 3.1 to 3.4 178

B.4 Calculation of Power Values in Table 5.1 185

List of Tables

2.1 *Critical Values c_α for the* LRT *with $C = 1, S = 2$ and unknown σ^2.* 62

2.2 *Power Comparison of the* LRT *with $C = 1, S = 2$ and unknown σ^2 against the Separate Test Approach, for $H : \boldsymbol{\theta} = (.3, 50, 25)^\top$ based on 10,000 Monte Carlo samples of sizes $N = 15$ and 25.* 63

2.3 *Power Comparison of the* LRT *with $C = 1, S = 2$ and unknown σ^2 against the Separate Test Approach, for $H : \boldsymbol{\theta} = (.5, 50, 45)^\top$ based on 10,000 Monte Carlo samples of sizes $N = 15$ and 25.* 64

3.1 *Maximum Pairwise Likelihood Estimates based on 50 Random Samples of Size $N = 50$ from the Grouped Continuous Model with $Q = 3$ and Parameters given by Case* (I). 90

3.2 *Maximum Pairwise Likelihood Estimates based on 50 Random Samples of Size $N = 100$ from the Grouped Continuous Model with $Q = 3$ and Parameters given by Case* (I). 91

3.3 *Maximum Pairwise Likelihood Estimates based on 50 Random Samples of Size $N = 50$ from the Grouped Continuous Model with $Q = 3$ and Parameters given by Case* (II). 92

3.4 *Maximum Pairwise Likelihood Estimates based on 50 Random Samples of Size $N = 100$ from the Grouped Continuous Model with $Q = 3$ and Parameters given by Case* (II). 93

4.1 *Three-Dimensional Array for the Appendicitis Data (Koepsel et al., 1981).* 128

4.2 *Maximum Likelihood Estimates of Parameters of General Mixed-Data Model for the Appendicitis Data.* 129

5.1 *Empirical Size and Power of χ^2 Test in Theorem 5.2 for $C = L = Q = 1$ and $S = 2$ based on 1,000 Monte Carlo Samples.* . 145
5.2 *Three-Dimensional Array for the Appendicitis Data (Koepsel et al., 1981) classified by Sex.* 146
5.3 *Maximum Likelihood Estimates of Parameters of General Mixed-Data Model for the Appendicitis Data classified by Sex.* 147

List of Figures

2.1 *Plots of the Power Function of* LRT *with* $C = 1, S = 2$ *and unknown* σ^2*, for* $p_0 = 0.5$ *and fixed State Means.* 60

2.2 *Contour Plots of the Power Function of* LRT *with* $C = 1, S = 2$ *and unknown* σ^2*, for* $N = 25$ *and* $p = 0.5$. 61

4.1 *Taxonomy of Models and Analytical Approaches in Mixed Multivariate Data Analysis.* 126

4.2 *Two Levels of Data Layout for the General Mixed-Data Model with* $Q = S = 2, L_1 = L_2 = 1$ *and* $C \geq 1$ 127

List of Notations

$\boldsymbol{\alpha}$	a vector of thresholds $(= (\alpha_q^{\ell_q}, \ell_q = 1, \cdots, L_q; q = 1, \cdots, Q)^\top)$
α	the level of significance of a test
$\alpha_q^{\ell_q}$	the ℓ_qth threshold for the qth ordinal variable Z_q, $\ell_q = 1, \cdots, L_q$; $q = 1, \cdots, Q$
$\mathbf{B}$	a $Q \times C$ matrix of regression parameters $(= (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_Q)^\top)$
$\boldsymbol{\beta}$	a stacked vector of regression coefficients $(= \text{vec}(\mathbf{B}))$
$\boldsymbol{\beta}_q$	the qth vector of regression coefficients for the conditional grouped continuous and general mixed-data models, $q = 1, \cdots, Q$
C	the total number of continuous variables
χ^2_{df}	a chi-square random variable with df degrees of freedom
$\chi^2_{df,\Delta}$	a noncentral chi-square random variable with df degrees of freedom and noncentrality parameter Δ
$\mathbf{D}$	the diagonal matrix of (conditional) standard deviations of $\mathbf{y}$
D	the total number of nominal categorical variables
$\Delta_{g'g''}$	the Kullback-Leibler divergence between populations $\mathcal{P}^{(g')}$ and $\mathcal{P}^{(g'')}$
$\mathbf{d}_s$	the vector equal to $\overline{\mathbf{y}}_s - \boldsymbol{\mu}_s$, $s = 1, \cdots, S$
F_{df_1,df_2}	an F random variable with degrees of freedom df_1 and df_2
$F^{\Delta}_{df_1,df_2}$	a noncentral F random variable with degrees of freedom df_1 and df_2 with noncentrality parameter Δ
$\phi_h(\cdot \mid \boldsymbol{\Sigma})$	the h-dimensional multivariate normal density function with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$
$\Phi_h(\cdot \mid \boldsymbol{\Sigma})$	the h-dimensional multivariate normal distribution function with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$
G	the total number of mixed-variate populations
$\boldsymbol{\Gamma}$	the covariance matrix of $\mathbf{y}$ and $\mathbf{y}^*$
$\boldsymbol{\gamma}$	a vector of standardized thresholds $(= (\gamma_q^{\ell_q}, \ell_q = 1, \cdots, L_q; q = 1, \cdots, Q)^\top)$
$\gamma_q^{\ell_q}$	standardized $\alpha_q^{\ell_q}$ equal to $\alpha_q^{\ell_q}/d_q$

$GLM(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$	the general location model with parameters $\boldsymbol{\pi}, \boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$
H	the null hypothesis
$\mathbf{H}$	the between-groups SSP matrix
$\mathbf{H}(\boldsymbol{\theta})$	the Hessian matrix with respect to $\boldsymbol{\theta}$
$\mathbf{I}_h$	the $h \times h$ identity matrix
$\mathcal{I}_P(\boldsymbol{\theta})$	the $P \times P$ Fisher information matrix with respect to $\boldsymbol{\theta}$
$\mathbf{J}_P$	the $P \times P$ pairwise information matrix
K	the alternative hypothesis
$\mathbf{K}_P$	the $P \times P$ matrix equal to $\sum_{i=1}^{N} \mathrm{E}_{\boldsymbol{\theta}} \left[(\partial p\ell_i / \partial \boldsymbol{\theta})(\partial p\ell_i / \partial \boldsymbol{\theta})^\top \right]$
$\mathcal{L}$	the likelihood function
$\xrightarrow{\mathcal{L}}$	denotes 'convergence in distribution'
$\boldsymbol{\ell}$	a possible value for the ordinal vector $\mathbf{z}$ $(= (\ell_1, \cdots, \ell_Q)^\top)$
ℓ_q	a possible value for Z_q $(= 1, \cdots, L_q)$
$\ell^{(C)}$	the log-likelihood function of the continuous data
$\ell(\boldsymbol{\theta})$	the log-likelihood function of $\boldsymbol{\theta}$
L_q	the total number of ordinal scores for Z_q, $q = 1, \cdots, Q$
λ	the likelihood ratio test statistic
$\Lambda_{df_1, df_2, df_3}$	a random variable with Wilks' λ-distribution with degrees of freedom df_1, df_2, and df_3
$\mathbf{M}^{-1}$	the inverse of the matrix $\mathbf{M}$
$\mathbf{M}^\top$	the transpose of the matrix $\mathbf{M}$
$\mathbf{M} \otimes \mathbf{N}$	the Kronecker product of matrices $\mathbf{M}$ and $\mathbf{N}$
$\boldsymbol{\mu}$	the $CS \times 1$ stacked vector of state means $(= (\boldsymbol{\mu}_1^\top, \cdots, \boldsymbol{\mu}_S^\top)^\top)$
$\boldsymbol{\mu}_s$	the sth state mean of $\mathbf{y}$, $s = 1, \cdots, S$
$\boldsymbol{\mu}_s^*$	the sth state mean of $\mathbf{y}^*$, $s = 1, \cdots, S$
$\boldsymbol{\mu}_{gs}$	the sth state mean for population $g = 1, \cdots, G$
N	the total number of observations $(= \sum_{s=1}^{S} n_s)$
n_s	the total number of observations belonging to state $s = 1, \cdots, S$
n_{gs}	the total number of observations belonging to state s in population $g = 1, \cdots, G$
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathbf{0}$	the zero matrix
p	the probability of $\mathbf{x} = \mathbf{x}_{(1)}$ for the general location model with $C = 1, S = 2$

P	the total number of independent parameters in the model
$\xrightarrow{p}$	denotes 'convergence in probability'
$\boldsymbol{\pi}$	the $S \times 1$ vector of state probabilities $(= (\pi_1, \cdots, \pi_S)^\top)$
$\boldsymbol{\pi}_g$	the $S \times 1$ vector of state probabilities for population g
π_s	the probability corresponding to the sth state $(\sum_{s=1}^{S} \pi_s = 1)$
$p\ell(\boldsymbol{\theta})$	the pairwise log-likelihood function
q	the probability of $\mathbf{x} = \mathbf{x}_{(1)}$ for the general location model with $C = 1, S = 2$ $(= 1 - p)$
Q	the total number of ordinal variables
$\mathbf{R}$	the $Q \times Q$ matrix of conditional polychoric correlations, given $\mathbf{y}$ (and $\mathbf{x}_{(s)}$)
$\mathbf{R}^*$	the $Q \times Q$ matrix of polychoric correlations
$r_{qq'}$	the (conditional) polychoric correlation between Z_q and $Z_{q'}$
$\rho_{\mathbf{x},Y}$	the polyserial correlation between $\mathbf{x}$ and Y
S	the total number of states
$\mathbf{S}$	the sample covariance matrix (uncorrected for bias)
s_d	the number of categories for the dth categorical variable U_d
$\mathbf{s}(\boldsymbol{\theta})$	the score vector with respect to $\boldsymbol{\theta}$
$\mathbf{s}_{PL}(\boldsymbol{\theta})$	the pairwise score vector with respect to $\boldsymbol{\theta}$
$\mathbf{S}_{gs}$	the sample covariance matrix (uncorrected for bias) of the continuous observations belonging to state s in population g
$\mathbf{S}_{pooled}$	the pooled sample covariance matrix (uncorrected for bias)
$\boldsymbol{\Sigma}$	the common covariance matrix of the continuous data
$\boldsymbol{\Sigma}^*$	the common covariance matrix of the latent variables
σ^2	the variance of Y in the general location model with $C = 1, S = 2$
$\boldsymbol{\tau}$	a $QS \times 1$ vector of state-specific effects $(= (\boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_S)^\top)$
$\boldsymbol{\tau}_s$	a $Q \times 1$ vector of effects of state s on $\mathbf{z}$ $(= \mathbf{D}^{-1}\boldsymbol{\mu}_s^* - \mathbf{B}\boldsymbol{\mu}_s)$
τ_{sq}	the effect of state s on the qth ordinal variable Z_q
Θ	the parameter space
$\boldsymbol{\theta}$	the location parameter of the general location model $(= (\boldsymbol{\pi}^\top, \boldsymbol{\mu}^\top)^\top)$; also the unknown parameter of a model
$\boldsymbol{\theta}_g$	the unknown parameter for population g
$\widehat{\boldsymbol{\theta}}$	the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$
$\widehat{\boldsymbol{\theta}}^{PL}$	the maximum pairwise likelihood (MPL) estimator of $\boldsymbol{\theta}$
$\text{tr}(\mathbf{M})$	the trace of the matrix $\mathbf{M}$

$\mathbf{u}$	a $D \times 1$ vector of nominal categorical variables $(= (U_1, \cdots, U_D)^\top)$
U_d	the dth categorical variable, $d = 1, \cdots, D$
$U^{(M)}$	Hotelling's generalized T^2 statistic
$\boldsymbol{\Upsilon}$	the asymptotic covariance matrix of vech($\mathbf{S}$)
$\mathbf{V}$	the asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}^{PL}$
vec($\mathbf{M}$)	the vector obtained by stacking the rows of $\mathbf{M}$
vech($\mathbf{M}$)	the vector containing the upper diagonal elements of $\mathbf{M}$
$\mathcal{W}_h(\boldsymbol{\Sigma}, df)$	the h-dimensional Wishart distribution with scale matrix $\boldsymbol{\Sigma}$ and degrees of freedom df
$\mathbf{x}$	a $S \times 1$ vector of binary variables such that $\sum_{s=1}^{S} X_s = 1$ $(= (X_1, \cdots, X_S)^\top)$
$\mathbf{x}_{(s)}$	the vector $\mathbf{x}$ with $X_s = 1$
$\mathbf{y}$	a $C \times 1$ vector of continuous variables $(= (Y_1, \cdots, Y_C)^\top)$
$\mathbf{y}^*$	a $Q \times 1$ vector of latent variables $(= (Y_1^*, \cdots, Y_Q^*)^\top)$
$\overline{\mathbf{y}}_s$	the sample mean vector of the observations belonging to state s
$\overline{\mathbf{y}}_{gs}$	the sample mean vector of the observations belonging to state s in population g
$\mathbf{z}$	a $Q \times 1$ vector of ordinal variables $(= (Z_1, \cdots, Z_Q)^\top)$

Chapter 1

Introduction

1.1 Background of the Thesis

Multivariate data containing mixtures of quantitative and qualitative variables arise frequently in practice. Catalano (1997) gives an example from developmental toxicology where fetal data from laboratory animals include binary, ordered categorical and continuous outcomes. Schafer (1997) describes a data set consisting of a variety of variable types used to investigate the validity of the Foreign Language Attitude Scale (FLAS), an instrument for predicting success in the study of foreign languages.

A number of simple, albeit *ad-hoc*, approaches to the analysis of such data have been used in applications. If, for example, the qualitative variables can be subjected to some scoring scheme, then all the variables can be treated as quantitative. On the other hand, all the variables can be treated as qualitative if the quantitative variables can be categorized through some grouping criteria. Another approach would be to analyze the quantitative and qualitative variables separately, and then to synthesize the two sets of results. However, as Krzanowski (1983) states, *"all these options involve some element*

of subjectivity, with possible loss of information, and do not appear very satisfactory in general." The first approach introduces considerable subjectivity in the numerical scoring scheme adopted and the second results in information loss due to categorization of the quantitative variables, while the third ignores any associations existing between the quantitative and qualitative variables.

The ideal general approach is to first specify a model for the joint distribution of the quantitative and qualitative variables, then to fit the model to the data at hand, and finally to use the parameter estimates to draw inferences. One way to specify the joint distribution of a number of variables is to express it as the product of the conditional distribution of a subset of the variables multiplied by the marginal distribution of the remaining variables. This suggests two routes that can be taken to formulate the joint distribution in the mixed case: (1) specify the marginal distribution of the qualitative variables and the conditional distribution of the quantitative variables, given the qualitative variables, or (2) specify the marginal distribution of the quantitative variables and the conditional distribution of the qualitative variables, given the quantitative variables.

The second approach was first mentioned by Cox (1972), who suggested that the joint distribution of a mixture of binary and continuous variables could be written as a logistic conditional distribution for the binary variables given the continuous variables multiplied by a marginal multivariate normal distribution for the latter. Cox and Wermuth (1992) pursued this idea further and pointed out its connection to probit-style and latent variable models. Such

models are now known as *conditional Gaussian regression models.* Recent works by Catalano and Ryan (1992), Moustaki (1996) and Sammel *et al.* (1997) have since generalized the model in several directions.

The first approach has received much attention in the literature in the context of the analysis of data with mixtures of categorical and continuous variables. Here it is assumed that the continuous variables have a different multivariate normal distribution for each possible setting of the categorical variable values, while the categorical variables have an arbitrary marginal multinomial distribution. This model has been termed the *conditional Gaussian distribution* (CGD), and it forms the central plank of graphical association models for the analysis of mixed categorical and continuous variables (Edwards, 1995; Whittaker, 1990; Lauritzen and Wermuth, 1989).

In the subsequent section, a number of issues that remain to be resolved concerning the analysis of data with mixtures of variable types are discussed. These issues concern the specification of models for mixed data as well as the ensuing inference, both estimation and tests of hypotheses, based on such models. The chapter concludes with a brief description and overview of the thesis.

1.2 Issues in the Analysis of Mixed Data

Despite the attention that mixed data analysis has recently received in the literature, a number of important methodological issues still remain to be addressed. These issues are identified and grouped into four general areas.

These are as follows.

(a) Construction of global tests of hypotheses on parameters of models for mixed data, with particular focus on the so-called *location hypotheses* that arise from Olkin and Tate's (1961) *general location model.*

(b) Implementation of computationally feasible methods of estimation for the *grouped continuous model* (Anderson and Pemberton, 1985), a latent variable model for multivariate ordinal data, and its extensions.

(c) Analysis of the most general case of mixed data consisting of nominal, ordinal and quantitative variables, a situation that, although commonplace in practice, has not been adequately treated in the literature.

(d) Extensions of conventional multivariate methods of calculating a distance measure to mixed-variable data settings.

1.2.1 Global Testing of Mixed Data Hypotheses

The models for mixed data mentioned in the previous section were originally developed as a device for testing hypotheses of independence or conditional independence. Lauritzen and Wermuth (1989) provided an all-encompassing treatment of multivariate dependencies in mixed data with the introduction of *conditional Gaussian* (CG) families.

Denote D categorical and C continuous variables as $\mathbf{u} = (U_1, \cdots, U_D)^\top$ and $\mathbf{y} = (Y_1, \cdots, Y_C)^\top$. Suppose that the dth categorical variable U_d has s_d categories, so that there are a total of $S = \prod_{d=1}^{D} s_d$ possible patterns of discrete

response, or states, for $\mathbf{u}$. A *full* CGD for $(\mathbf{u}^\top, \mathbf{y}^\top)^\top$ assumes that the joint probability density of observing state s of $\mathbf{u}$ and $\mathbf{y}$ is

$$\pi_s \times (2\pi)^{-C/2} \mid \boldsymbol{\Sigma}_s \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_s)^\top \boldsymbol{\Sigma}_s^{-1}(\mathbf{y}-\boldsymbol{\mu}_s)\right\}. \tag{1.1}$$

That is, it assumes that if $\mathbf{u}$ falls in the sth state (or discrete response pattern), then $\mathbf{y} \sim \mathcal{N}_C(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$, while the probability that $\mathbf{u}$ falls in state s is π_s ($\sum_{s=1}^{S} \pi_s = 1$). The density in (1.1) can be rewritten in the form

$$\exp\left\{\phi_s + \boldsymbol{\psi}^\top \mathbf{y} - \frac{1}{2}\mathbf{y}^\top \boldsymbol{\Sigma}_s \mathbf{y}\right\}. \tag{1.2}$$

The triple $(\pi_s, \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$ comprising, respectively, the sth state probability, the sth state mean and the sth state dispersion matrix, in (1.1) are called the *moment* parameters of the CGD, while the parameters in (1.2) are the *canonical* parameters of the CGD. Here the ϕ_s are discrete canonical parameters and the $\boldsymbol{\psi}_s$ are $C \times 1$ vectors of linear canonical parameters. Technical aspects concerned with fitting these models and likelihood-based estimation and hypothesis-testing are covered in the references cited earlier.

Although this thesis is not concerned specifically with graphical modelling, it is pertinent to note that the CGD model has appeared previously in the literature in various contexts. Moustafa (1957) was the first to consider the full CGD model in the analysis of multi-way tables. Another CGD model was introduced by Olkin and Tate (1961) for mixed binary and continuous data, and has since been known as the *general location model* (Schafer, 1997; Little and Rubin, 1987; Little and Schluchter, 1985). This particular model assumes a uniform dispersion matrix $\boldsymbol{\Sigma}$ across the states and is called a *homogeneous*

CGD in the graphical modelling literature. Olkin and Tate (1961), while considering canonical correlations between the binary and continuous variables, established results connecting these canonical correlations and the state means, and investigated the distribution theory of their estimators. Krzanowski (1983) used the general location model in the calculation of distance between two populations with mixtures of binary and continuous variables, and applied the resulting distance in discriminant analysis.

One particular aspect of mixed data inference that has received little attention so far are the so-called *location hypotheses*, for which the construction of reasonable statistical tests remains an important and so far unaddressed problem in such applications as quality control (de Leon and Carrière, 2000) and clinical studies (Afifi and Elashoff, 1969). The problem of interest is to test

$$H : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad \text{against} \qquad K : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \tag{1.3}$$

for some specified $\boldsymbol{\theta}_0$, where $\boldsymbol{\theta}^\top = (\boldsymbol{\pi}^\top, \boldsymbol{\mu}^\top)$, with $\boldsymbol{\mu}^\top = (\boldsymbol{\mu}_1^\top, \cdots, \boldsymbol{\mu}_S^\top)$ the $CS \times 1$ vector of state means. Hypothesis H in (1.3) is referred to in the literature as the *one-sample location hypothesis*, and much work has been done for the case with continuous data. Afifi and Elashoff (1969) tackled the two-sample mixed data problem and obtained two tests, one based on the Kullback-Leibler divergence (Kullback, 1968, pp. 6-7) and another on the likelihood ratio approach.

The simple hypothesis in (1.3) is of particular interest in such applications as quality control charting situations, where a process is monitored

through time via *control charts* to evaluate whether the process is performing according to certain industrial specifications (see, e.g., Montgomery, 1985). Here, the control limits of a multivariate control chart are to be set up to simultaneously and more effectively chart both the discrete and continuous characteristics, as opposed to charting them separately with univariate control charts. In this context, the alternative hypothesis may correspond to a signal for the process being *out of control*. The absence of a signal in the multivariate chart precludes the presence of signals in the univariate charts.

In practice, the analytic strategy with mixed data has been to perform tests on the parameters separately. This approach entails the problem of multiple significance testing, to which the simplest solution is to adjust the level of each test to control the overall level. Such an approach may lose power quite substantially because the correlations between the variables are not utilized explicitly in constructing the test statistic (Pocock *et al.*, 1987). An alternative approach is to treat the problem in a multivariate setting to come up with a single test based on all the variables. O'Brien (1984) and Pocock *et al.* (1987) studied one such global test statistic in the context of clinical trials.

The likelihood ratio approach is of central importance in the construction of global tests of location hypotheses for mixed data. This approach allows the problem to be treated from a multivariate perspective to simultaneously test both the discrete and continuous parameters of the general location model. In addition to the hypothesis (1.3), its natural extension to the multi-sample

situation is also of interest, which involves testing whether the respective parameters $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_G$ of G mixed-variate populations, distributed according to the general location model, are all equal. That is, it is desired to construct a test of

$$H : \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_G \qquad \text{against} \qquad K : \boldsymbol{\theta}_g \neq \boldsymbol{\theta}_{g'}, \tag{1.4}$$

for some $g \neq g'$. Note that (1.4) generalizes the classical *multivariate analysis of variance* (MANOVA) problem to the mixed data setting. Despite extensive results available on the classical MANOVA problem, they could not be readily applied to the mixed data case. Besides the earlier works of Moustafa (1957), Ogawa *et al.* (1957), Afifi and Elashoff (1969), Morales *et al.* (1998), and de Leon and Carrière (2000) recently, this problem remains and needs to be addressed further.

1.2.2 Computationally Efficient Estimation Methods

Many authors have considered the analysis of multivariate ordinal data, especially as they occur in many studies in the social sciences. Although no consensus exists about the manner by which the analysis should proceed, one of the more common approaches has been to postulate the existence of continuous latent variables underlying the observed data, and to assume that these latent variables follow some continuous multivariate distribution, a partitioning of which gives rise to the levels of the (observed) ordinal data. Models for ordinal data specified this way were first suggested by Pearson (1904), and they have been further developed over the years (Anderson and Philips, 1981; Mc-

Cullagh, 1980).

One such model, called the *grouped continuous model* (Anderson and Pemberton, 1985), considers the multivariate normal distribution as the distribution for the latent variables. In it, an ordinal vector $\mathbf{z} = (Z_1, \cdots, Z_Q)^\top$ is observed, where Z_q has $1 < \cdots < L_q$ ordered levels, $q = 1, \cdots, Q$, and corresponding to $\mathbf{z}$ is a vector of unobservable continuous latent variables $\mathbf{y}^* = (Y_1^*, \cdots, Y_Q^*)^\top$, distributed according to the multivariate normal distribution $\mathcal{N}_Q(\mathbf{0}, \mathbf{R}^*)$ with mean vector $\mathbf{0}$ and correlation matrix $\mathbf{R}^*$, such that $Z_q = \ell_q$ if and only if $\alpha_q^{\ell_q - 1} < Y_q^* \leq \alpha_q^{\ell_q}$, $\ell_q = 1, \cdots, L_q$, with $\{\alpha_q^0 = -\infty < \alpha_q^1 < \cdots < \alpha_q^{L_q} < \alpha_q^{L_q+1} = +\infty\}$ the unknown *cutpoints* or *thresholds* for Z_q, $q = 1, \cdots, Q$. The correlations in $\mathbf{R}^*$ are usually called *polychoric correlations*. This model is a generalization of the univariate grouped continuous model discussed earlier by Anderson and Philips (1981) and McCullagh (1980), and is closely linked to *probit models* in latent variable theory.

Maximum likelihood estimation for the grouped continuous model in the bivariate case (i.e., $Q = 2$) has been considered by a number of authors in the past (see, e.g., Drasgow, 1986; Olsson, 1979). Although the extension to higher dimensions is straightforward (see, e.g., Lee, 1985; Lee *et al.*, 1989), likelihood estimation is computationally impractical, as it involves the evaluation of high-dimensional normal integrals, which require a large amount of time to evaluate especially when the dimension is high. A computationally more efficient approach is thus desired.

Several alternative estimation methods that all rely on partitioning the

model into its sub-models have been proposed in the literature. Anderson and Pemberton (1985) developed a computationally feasible two-step approach that consists in first estimating the thresholds marginally, and then estimating the polychoric correlations by maximizing the pairwise marginal likelihoods with the thresholds replaced by their estimates. Lee and Poon (1987) and Lee and Lau (1986) studied the generalized least squares method and a two-step variant, and compared them with the maximum likelihood approach. Poon *et al.* (1990) applied the *partition maximum likelihood* (PML) method (Poon and Lee, 1987; 1986) in the multi-sample case. The method partitions the model into its univariate and bivariate sub-models, and estimates the parameters from these sub-models, then averages them in the end to obtain the final estimates. Bedrick *et al.* (2000) recently modified the method by working exclusively with the bivariate sub-models.

The appeal of the above methods lies in the fact that the computational burden of the maximum likelihood method is reduced to a considerable extent. As well, limited simulation studies (Poon *et al.*, 1990; Lee and Poon, 1987) appear to show that they are comparable with maximum likelihood estimation. However, concern remains regarding the efficiency of the estimates. By estimating parameters for individual Z_q's and pairs $\{Z_q, Z_{q'}\}$ separately, the ordinal variables are treated as though they are independent. Aside from yielding multiple sets of estimates with no clear criterion for combining them, the efficiency of the final estimates may be compromised. A simultaneous estimation of the parameters that is not as computationally expensive as the

maximum likelihood method is preferred, as it yields a single set of estimates and may lead to significant gains in efficiency.

The extension of the grouped continuous model to the case of mixed ordinal and quantitative data has been studied by Anderson and Pemberton (1985), Poon and Lee (1987; 1986). This is accomplished by assuming that the continuous variables share a joint multivariate normal distribution with the latent variables, and the thresholds and polychoric correlations are defined in terms of the conditional distribution of the latent variables (or the ordinal data) given the continuous data. In addition to these parameters, additional parameters representing the *polyserial correlations*, or the correlations between the ordinal and continuous variables, are introduced in the model. Poon and Lee (1992), Poon *et al.* (1990), and Lee *et al.* (1989) also extended the model to the analysis of several independent samples. In all this work, maximum likelihood and PML became the basis of the estimation methods used. As in the single-sample case, computationally more efficient alternatives to maximum likelihood that do not suffer from the same shortcomings as do the partition methods need to be explored in the context of multi-sample analysis.

1.2.3 Models for Mixed Nominal, Ordinal and Continuous Data

The most general case of mixed data encountered in practice are those which include mixtures of nominal, ordinal and continuous variables. Suppose that, in addition to the vectors $\mathbf{u}_{D\times 1}$ and $\mathbf{y}_{C\times 1}$ of nominal and continuous variables, a vector $\mathbf{z}_{Q\times 1}$ of ordinal variables is observed. One way to go about the

analysis of such data is to use the grouped continuous model in modelling the joint distribution of $\mathbf{u}$, $\mathbf{y}$ and $\mathbf{z}$, thus implicitly assuming an underlying latent variable structure for the nominal variables. Although this approach has been previously used for dichotomous nominal variables (see, e.g., Bock, 1972), it is, in general, inappropriate in the polytomous nominal case.

A better alternative is to model the joint distribution of $\mathbf{u}$, $\mathbf{y}$ and $\mathbf{z}$ by the general location model by considering the $(D+Q)\times 1$ vector $(\mathbf{u}^\top, \mathbf{z}^\top)^\top$ as a categorical vector with $S \times \prod_{q=1}^{Q} L_q$ states. Little and Schluchter (1985) have used this approach for the St. Louis Risk Research Project data. It has also been used in producing multiply-imputed public-use data files (Schafer, 1997; Rubin, 1996). While the general location model may, in principle, be used in this case, it may be inadequate, and hence, inappropriate, for two reasons. Firstly, there is no clear-cut manner of accounting for ordinal information, and secondly, there is no explicit way of incorporating correlations between the nominal vector $\mathbf{u}$ and the ordinal vector $\mathbf{z}$, and between the ordinal vector $\mathbf{z}$ and the continuous vector $\mathbf{y}$.

There is therefore a need to define appropriate models for mixed data with ordinal variables that can better describe the interrelationships between the different variable types as well as better account for the information arising from the variables' different levels of measurement. A possible approach, and the one adopted in this thesis, is a compromise between the two approaches mentioned above. That is, the joint distribution of $\mathbf{u}$, $\mathbf{y}$ and $\mathbf{z}$ is specified by breaking it up in terms of a general location model for $\mathbf{u}$ and $\mathbf{y}$, and a

conditional grouped continuous model for $\mathbf{z}$, conditional on $\mathbf{u}$ and $\mathbf{y}$. This approach accounts for the ordinal information in the ordinal variables as well as incorporates, both explicitly or implicitly, correlations among the three groups of variables. As well, such formulation of the joint distribution necessarily assumes a hierarchical structure, which allows for easier incorporation of covariate effects when extended to the regression setting. Note also that the approach basically unifies the two models for mixed data that have been studied previously in the literature.

The approach, however, gives rise to a number of problems related to model inference. Estimation of the model parameters becomes more involved, especially for those associated with the conditional grouped continuous model, as it involves regression (representing polyserial correlations between $\mathbf{y}$ and $\mathbf{z}$) and state-effect parameters induced by $\mathbf{y}$ and $\mathbf{u}$, respectively, in addition to the polychoric correlations and threshold parameters. Maximum likelihood estimation is possible for this model, although it may be impractical. A simpler, more computationally feasible alternative is needed for estimating the ordinal data model parameters.

Tests of hypotheses concerning comparisons of state means, the polychoric and polyserial parameters among the ordinal and continuous variables across and within states also need to be constructed.

1.2.4 Multivariate Methods for Mixed Data

Methods for the analysis of multivariate continuous data are well documented (e.g., Seber, 1984; Mardia *et al.*, 1979). So are those for multivariate discrete

data (Bishop *et al.*, 1975). Methods for multivariate mixed data, however, are not as well developed as the two cases. In view of the ubiquitousness of data with mixed variable types in practice, appropriate analytical methods are needed that will account for the different measurement levels and associations in the data efficiently and effectively. In particular, there is the need to define a distance measure among groups that may be used for mixed data.

Bedrick *et al.* (2000) and Lapidus (1998) recently studied methods for the analysis of mixed continuous and ordinal data. Specifically, they proposed a generalization of the *Mahalanobis distance* (Mardia *et al.*, 1979, p. 31) to populations of mixed continuous and ordinal variables and used the generalized distance in discriminant analysis. Bedrick *et al.* (2000) and Lapidus (1998) used the conditional grouped continuous model to model the joint distribution of the variables, and showed that the Mahalanobis distance can be decomposed as a sum of two components, one based on the continuous variables, and the other on the ordinal variables. Large-sample results on maximum likelihood estimators and tests of hypotheses were derived, as well as on a modified PML procedure. The generalized Mahalanobis distance follows along the same lines as that proposed by Bar-Hen and Daudin (1995). Bar-Hen and Daudin (1995) applied the *Kullback-Leibler divergence* (Kullback, 1968, pp. 6-7) to the general location model to derive a generalization of the Mahalanobis distance to data with mixed nominal and continuous data. Bar-Hen and Daudin (1995) and Daudin and Bar-Hen (1999) investigated its use in discriminant analysis, with the latter investigating variable selection in particular.

Another generalized distance for mixed-variate populations was previously studied by Krzanowski (1984). Instead of the Kullback-Leibler divergence, Krzanowski (1984) used *Matusita's distance* (Matusita, 1956) to derive a distance between populations consisting of nominal and continuous variables defined by the general location model. Nakanishi (1996) proposed another mixed-data distance that includes Bar-Hen and Daudin's (1995) and Krzanowski's (1984) distances. More recently, Bar-Hen and Daudin (1998) obtained the asymptotic distribution of Krzanowski's (1984) generalized Matusita's distance.

A direct extension of the above to mixed data with nominal, ordinal and continuous variables is possible by considering the approach described earlier for modelling the joint distribution of such mixed data. In this approach, the joint distribution of the nominal and continuous variables is modelled by the general location model while that of the ordinal variables, conditional on the nominal and continuous variables, is modelled by the grouped continuous model with regression effects due to the continuous variables and state-effects due to the nominal variables, in addition to the ordinal level-effects and the polychoric correlations. Such an approach for specifying the joint distribution accounts for the ordinal information in the data and the associations among the various types of variables.

1.3 Overview of the Thesis

The objective of this thesis is to address the issues raised in the previous section concerning the analysis of mixed-variable data.

Chapter 2 begins with a formal definition of the general location model and develops the framework for simultaneously testing its parameters. A likelihood ratio approach is adopted and several tests are obtained that allow for globally testing the discrete and continuous parameters of the general location model in the one-sample and multi-sample settings. The performance of the proposed one-sample tests is also compared, in terms of power and the ability to maintain the nominal level, with the one that carries out separate tests of the discrete and continuous parameters.

Chapter 3 gives a general introduction to the grouped continuous model for mixed continuous and ordinal data. This chapter briefly reviews conventional methods, including maximum likelihood, of parameter estimation proposed by researchers in the past, and proposes a new estimation method based on the *pairwise likelihood approach* (e.g., Kuk and Nott, 2000). The latter is less computationally demanding than the former, and is conceptually more appealing than the partition methods. An investigation, via simulation, of the efficiency and bias of the estimates obtained using *maximum pairwise likelihood estimation* is reported as well. This chapter also derives the asymptotic distribution of the maximum pairwise likelihood estimator and discusses related large-sample tests of hypotheses.

In Chapter 4, the *general mixed-data model* for mixed continuous, ordi-

nal and nominal data is developed and is linked to the general location and grouped continuous models. Besides maximum likelihood estimation, a more computationally efficient alternative based on the pairwise likelihood approach, which is an extension of the methods proposed in Chapter 3, is outlined for the model. The asymptotic distributions of the estimates are also obtained for use in the construction of large-sample tests of hypotheses.

Chapter 5 proposes a generalization of the Mahalanobis distance to mixed data with nominal, ordinal and continuous variables. Applying the Kullback-Leibler divergence to the general mixed-data model of Chapter 4, a generalization of the Mahalanobis distance is derived that further extends those previously proposed by Bedrick *et al.* (2000) and Lapidus (1998), and by Bar-Hen and Daudin (1995). Asymptotic properties are obtained for the generalized distance and the results of a simulation study on the performance of tests of hypotheses are reported. An example is also presented to illustrate its application.

Finally, a summary of the results of the thesis is provided in Chapter 6. Promising areas for future research are identified as well.

Chapter 2

Location Hypothesis Tests for Mixed Data

2.1 Introduction

In this chapter, a hypothesis-testing problem that arises with multivariate data having both continuous and discrete variables is studied. In the *one-sample* case, the problem concerns the construction of statistical tests for the null hypothesis that the *location parameters* are equal to some specified value. In the *multi-sample* case, it entails testing whether the location parameters are the same in two or more distinct populations. The former arises in quality control applications where a manufacturing process is monitored with respect to product characteristics which may involve both continuous and discrete variables. The latter is often encountered in medical and health studies when several treatments for some disease or disorder are compared in terms of outcomes that include both continuous and discrete characteristics of the patients.

In these applications, the hypotheses concerning the parameters of the mixed-variable data are tested separately by applying conventional methods

for discrete and continuous variables (see, e.g., Birdsall *et al.*, 1997). Alternatively, one can consider the use of *global tests* (Pocock *et al.*, 1987) based on an appropriate multivariate model for the data to compare the parameters of a mixed-variate population against some target value or the parameters of several (two or more) such populations. Global tests combine information from all the variables by fully exploiting the multivariate nature of the data thus resulting in increased power for the tests (de Leon and Carrière, 2000; Pocock *et al.*, 1987; O'Brien, 1984). Unfortunately, standard multivariate approaches do not directly apply, and suitable methods have not been widely studied.

The goal of this chapter is two-fold. First, global one-sample location tests for mixed multivariate data are constructed. This is accomplished by adopting the general location model for mixed continuous and discrete data and using the likelihood ratio approach to derive the tests. Second, these tests are generalized to the multi-sample setting and in so doing, previous work on multi-sample location tests are extended to allow for the case of data with mixed continuous and discrete variables.

Despite the recent interest on mixed data analysis, only a few papers have considered similar problems. Moustafa (1957) studied a multi-factor experiment where the response variables consist of continuous and discrete variables jointly distributed according to the full CGD. He considered hypotheses concerning the independence and conditional independence of the responses, and proceeded to construct asymptotic likelihood ratio tests, the theory for which was previously studied in Ogawa *et al.* (1957). A related problem was ad-

dressed by Olkin and Tate (1961), who introduced the general location model and derived tests of independence between continuous and discrete variables via canonical correlation theory.

Afifi and Elashoff (1969) were the first to address location hypothesis-testing in mixed data situations. They considered the problem in the two-sample case, for which they derived likelihood ratio and information-theoretic tests. The exact sampling distributions of the test statistics were also obtained; however, no comparison was undertaken of their small-sample performance. Additionally, they showed that the Hotelling T^2 statistic (Mardia *et al.*, 1979, pp. 76-77) is not consistent against the location hypothesis considered.

A more recent paper by Morales *et al.* (1998) introduced a general class of dissimilarity or entropy-type measures to obtain test statistics for various hypotheses, including those considered here, involving mixed continuous and categorical data and used the asymptotic theory of these statistics to construct the tests.

This chapter introduces the general notation and framework for modelling and testing with mixed multivariate data. The general location model, which is adopted as the model for the mixed data, is formally defined in § 2.2. The one-sample problem for both the known and unknown covariance matrix cases is then discussed in § 2.3. While a general multivariate situation is discussed, a detailed study of the bivariate case is provided. The multi-sample problem is investigated in § 2.4. The chapter concludes with a discussion in § 2.5 of the performance of the tests and other related issues.

2.2 General Location Model

Definition 2.1 *Suppose* $\mathbf{x} = (X_1, \cdots, X_S)^\top$ *and* $\mathbf{y} = (Y_1, \cdots, Y_C)^\top$ *are vectors of binary and continuous variables, respectively, such that* X_s *is either 1 or 0 and* $\sum_{s=1}^{S} X_s = 1$. $(\mathbf{x}^\top, \mathbf{y}^\top)^\top$ *is said to be distributed according to the general location model if and only if* $\mathbf{x}$ *has a multinomial distribution given by*

$$p(\mathbf{x}) = \prod_{s=1}^{S} \pi_s^{x_s},$$

and $\mathbf{y}$ *has, given* $X_s = 1$, *a multivariate normal distribution with mean* $\boldsymbol{\mu}_s = (\mu_{1s}, \cdots, \mu_{Cs})^\top$ *and covariance matrix* $\boldsymbol{\Sigma}$. *The model is denoted by* $GLM(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, *where* $\boldsymbol{\pi}^\top = (\pi_1, \cdots, \pi_S)$ *is the vector of state probabilities and* $\boldsymbol{\mu}^\top = (\boldsymbol{\mu}_1^\top, \cdots, \boldsymbol{\mu}_S^\top)$ *is* $CS \times 1$ *vector of state means. Here,* $\pi_s = \Pr(X_s = 1) > 0$ *and* $\sum_{s=1}^{S} \pi_s = 1$.

The above definition was originally given by Olkin and Tate (1961). Note, however, that Olkin and Tate (1961) referred to the above model as the *location model.* The term *general location model* was first used by Krzanowski (1980) to refer to its extension allowing for general categorical—not necessarily dichotomous—discrete variables. As these variables can always be defined in terms of binary variables, no distinction between the two terminologies shall be made. Further remarks on the model are given below.

Remark 2.2.1 *The model above arises for a vector* $\mathbf{u} = (U_1, \cdots, U_D)^\top$ *consisting of categorical variables where the* d*th variable* U_d *has* s_d *categories, so that there are a total of* $S = \prod_{d=1}^{D} s_d$ *possible states for* $\mathbf{u}$. *In this case, we can define* $\mathbf{x}$ *as* $X_s = 1$ *if* $\mathbf{u}$ *falls in state* s *and 0 otherwise, and* $\sum_{s=1}^{S} X_s = 1$.

For notational convenience, the vector $\mathbf{x}$ *for which* $X_s = 1$ *is denoted by* $\mathbf{x}_{(s)}$.

Remark 2.2.2 *Note that the discrete vector* $\mathbf{u}$ *defines an* $s_1 \times \cdots \times s_D$ *contingency table with* S *cells (or states). For each given state, the model assumes a multivariate normal distribution for* $\mathbf{y}$ *with mean* $\boldsymbol{\mu}_s$, $s = 1, \cdots, S$, *and covariance matrix* $\boldsymbol{\Sigma}$, *written as* $(\mathbf{y} \mid \mathbf{x}_{(s)}) \sim \mathcal{N}_C(\boldsymbol{\mu}_s, \boldsymbol{\Sigma})$.

Remark 2.2.3 *Observe that* $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_S$ *if and only if* $\mathbf{x}$ *and* $\mathbf{y}$ *are independent. See Olkin and Tate (1961) for details.*

Note that there are a total of $(S-1) + CS + C(C+1)/2$ independent parameters in the model. To reduce the number of parameters, it is suggested to impose log-linear restrictions on $\boldsymbol{\pi}$ and a (linear) hierarchical structure for $\boldsymbol{\mu}$ (Schafer, 1997; Raghunathan and Grizzle, 1995; Little and Rubin, 1987; Little and Schluchter, 1985).

Recent extensions of the general location model were given by Barnard *et al.* (2000), Liu and Rubin (1998) and Fitzmaurice and Laird (1997, 1995). The first two papers attempted to relax the homogeneity assumption in the model—the former by factorizing the covariance matrix in terms of correlations and standard deviations and allowing for the standard deviations to vary across states, and the latter by considering ellipsoidally elliptic distributions, including the multivariate t distribution, as alternative distributions for the continuous vector $\mathbf{y}$. Fitzmaurice and Laird (1997, 1995) generalized the model to the regression setting by allowing for auxiliary variables to be incorporated in the model for mixed binary and continuous outcomes.

2.2.1 Likelihood Function

Suppose $(\mathbf{x}_1^\top, \mathbf{y}_1^\top)^\top, \cdots, (\mathbf{x}_N^\top, \mathbf{y}_N^\top)^\top$ are a random sample from $GLM(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Without loss of generality, assume $\mathbf{x}_{N_j+1} = \cdots = \mathbf{x}_{N_{j+1}} = \mathbf{x}_{(j+1)}$ so that $\mathbf{y}_{N_j+1}, \cdots, \mathbf{y}_{N_{j+1}}$ are independently and identically distributed as $\mathcal{N}_C(\boldsymbol{\mu}_{j+1}, \boldsymbol{\Sigma})$, $j = 0, 1, \cdots, S-1$, where $N_0 = n_0 = 0$, $N_j = \sum_{s=0}^{j} n_s$ for $j = 1, \cdots, S-1$, $N_S = N = \sum_{s=1}^{S} n_s$, and n_s is the number of observations in state $s = 1, \cdots, S$. Then,

$$\mathcal{L} = \prod_{s=1}^{S} \frac{\pi_s^{n_s}}{|\ 2\pi\boldsymbol{\Sigma}\ |^{N/2}} \times \exp\left[-\frac{1}{2} \sum_{j=0}^{S-1} \sum_{i=N_j+1}^{N_{j+1}} (\mathbf{y}_i - \boldsymbol{\mu}_{j+1})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_{j+1}) \right], \quad (2.1)$$

as the likelihood function of the whole sample. Note that $\mathcal{L}$ consists of two parts, $\mathcal{L}^{(D)}$ and $\mathcal{L}^{(C)}$, the first corresponding to the usual multinomial sample likelihood and the second to that of a multivariate normal sample. Equivalently, the log-likelihood function can be written as follows (Mardia *et al.*, 1979, p. 97):

$$\ell = \sum_{s=1}^{S} n_s \log \pi_s - \frac{N}{2} \log |\ 2\pi\boldsymbol{\Sigma}\ | - \sum_{s=1}^{S} \frac{n_s}{2} \mathrm{tr}\left[\boldsymbol{\Sigma}^{-1} \left(\mathbf{S}_s + \mathbf{d}_s \mathbf{d}_s^\top \right) \right], \quad (2.2)$$

where $\mathrm{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$, $\mathbf{d}_s = \overline{\mathbf{y}}_s - \boldsymbol{\mu}_s$, $\overline{\mathbf{y}}_s$ and $\mathbf{S}_s$ are the sample mean and sample covariance matrix (uncorrected for bias), respectively, of the observations in state $s = 1, \cdots, S$.

Example 2.1 *Consider the simplest general location model where $S = 2$ and $C = 1$. Suppose $\mathbf{x} = \mathbf{x}_{(1)}$ and $\mathbf{x} = \mathbf{x}_{(2)}$ have respective probabilities p and $q = 1 - p$, and the conditional distributions of Y for $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$ are assumed to be $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$, respectively. Given a random sample of size*

N, the likelihood function in (2.1) is given by

$$\mathcal{L} = p^n q^{N-n} (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{Q(\mu_1, \mu_2)}{2\sigma^2}\right], \tag{2.3}$$

where $Q(\mu_1, \mu_2) = \sum_{i=1}^{n}(Y_i - \mu_1)^2 + \sum_{i=n+1}^{N}(Y_i - \mu_2)^2$, *and it is assumed that the first n observations have* $\mathbf{x} = \mathbf{x}_{(1)}$.

The model in Example 2.1 is studied in de Leon and Carrière (2000), where $\mathbf{x}$ is replaced by a Bernoulli random variable. It can be formulated as a latent variable model by supposing an unobservable continuous variable Y^* underlying the binary variable. This approach is pursued in the following example.

Example 2.2 *Assume* $(Y^*, Y)^\top$ *has a bivariate normal distribution with* $\mathrm{E}(Y) = \mu$, $\mathrm{E}(Y^*) = 0$, $\mathrm{var}(Y) = \tau^2$, $\mathrm{var}(Y^*) = 1$, *and* $\mathrm{cov}(Y^*, Y) = \rho\tau$, *such that* $\mathbf{x} = \mathbf{x}_{(1)}$ *whenever* $Y^* \leq \alpha$ *and* $\mathbf{x} = \mathbf{x}_{(2)}$ *otherwise. Since* $Y \mid Y^* \sim \mathcal{N}(\mu + \rho\tau Y^*, \tau^2(1-\rho^2))$ *and* $Y^* \sim \mathcal{N}(0,1)$, *then* $p = \Pr(\mathbf{x} = \mathbf{x}_{(1)}) = \Phi(\alpha)$, $\mu_1 = \mu + \rho\tau$, $\mu_2 = \mu$, *and* $\sigma^2 = \tau^2(1-\rho^2)$, *with* Φ *the standard normal distribution function. The same model in Example 2.1 is thus obtained by a reparametrization of the latent variable model, and the same likelihood function as in (2.3) is obtained.*

The model in Example 2.2 is studied in detail by Tate (1955); an extension is given by Hannan and Tate (1965).

2.2.2 Maximum Likelihood Estimation

Maximum likelihood estimators (MLEs) of unknown parameters of the model are obtained by maximizing either (2.1) or (2.2). Because the parameter space

is simply the product of the individual spaces of the discrete and continuous parameters, MLEs are obtained by maximizing $\mathcal{L}^{(D)}$ and $\mathcal{L}^{(C)}$ separately.

The MLEs of $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are easily found to be

$$\begin{aligned}
\widehat{\boldsymbol{\pi}}^{\top} &= \sum_{i=1}^{N} \mathbf{x}_i^{\top}/N \\
&= (n_1/N, \cdots, n_S/N);
\end{aligned}$$

$$\begin{aligned}
\widehat{\boldsymbol{\mu}}^{\top} &= (\sum_{i=1}^{n_1} \mathbf{y}_i^{\top}/n_1, \cdots, \sum_{i=N_{S-1}+1}^{N} \mathbf{y}_i^{\top}/n_S) \\
&= (\overline{\mathbf{y}}_1^{\top}, \cdots, \overline{\mathbf{y}}_S^{\top}) \\
&\equiv \overline{\mathbf{y}}^{\top}; \qquad (2.4)
\end{aligned}$$

$$\begin{aligned}
N\widehat{\boldsymbol{\Sigma}} &= \sum_{j=0}^{S-1} \sum_{i=N_j+1}^{N_{j+1}} (\mathbf{y}_i - \overline{\mathbf{y}}_{j+1})(\mathbf{y}_i - \overline{\mathbf{y}}_{j+1})^{\top} \\
&= \sum_{s=1}^{S} n_s \mathbf{S}_s/N \\
&\equiv N\mathbf{S}_{pooled},
\end{aligned}$$

using standard results on multinomial (Bishop *et al.*, 1975) and multivariate normal distributions (Mardia *et al.*, 1979). Here it is assumed that $n_s > 0$ $\forall s$; otherwise, parameters corresponding to states with zero counts become inestimable. Note that the MLE of $\boldsymbol{\mu}$ remains unchanged, regardless of whether $\boldsymbol{\Sigma}$ is known or not, or whether homogeneity is assumed or not. However, the same is not true of $\widehat{\boldsymbol{\Sigma}}$ when $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ is known. In this case, it can be shown that the MLE of $\boldsymbol{\Sigma}$ becomes $\widehat{\boldsymbol{\Sigma}}_0 = \sum_{j=0}^{S-1} \sum_{i=N_j+1}^{N_{j+1}} (\mathbf{y}_i - \boldsymbol{\mu}_{0,j+1})(\mathbf{y}_i - \boldsymbol{\mu}_{0,j+1})^{\top}/N$ (e.g., Seber, 1984, p. 67).

As detailed in Schafer (1997, pp. 334-337), the MLEs of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are also the *least-squares estimates* in the *standard multivariate regression* of $\mathbf{y}$ on $\mathbf{x}$.

Example 2.3 *Consider again Example 2.1. The* MLE*s are then given by* $\widehat{p} = n/N$ $(0 < n < N)$, $\widehat{\mu}_1 = \sum_{i=1}^{n} Y_i/n = \overline{Y}_1$, $\widehat{\mu}_2 = \sum_{i=n+1}^{N} Y_i/(N-n) = \overline{Y}_2$, *and* $\widehat{\sigma}^2 = Q(\overline{Y}_1, \overline{Y}_2)/N$. *These estimates are unchanged except when* μ_1 *and* μ_2 *are known, in which case the* MLE *of* σ^2 *becomes* $\widehat{\sigma}^2 = Q(\mu_1, \mu_2)/N$.

Property 2.1 $\mathrm{E}(\widehat{\boldsymbol{\pi}}) = \boldsymbol{\pi}$, $\mathrm{E}(\overline{\mathbf{y}}) = \boldsymbol{\mu}$ *and* $\mathrm{E}(\mathbf{S}_{pooled}) = (N-S)\boldsymbol{\Sigma}/N$.

Proof. The first two follow immediately from the unbiasedness of $\widehat{\boldsymbol{\pi}}$ and $\overline{\mathbf{y}}$ while the third is obtained from the fact that $\mathrm{E}(\mathbf{S}_s \mid n_s) = (n_s - 1)\boldsymbol{\Sigma}/n_s$ $\forall s$.

□

Property 2.2 *Given* $(n_1, \cdots, n_S)^\top$, $\overline{\mathbf{y}} \sim \mathcal{N}_{CS}(\boldsymbol{\mu}, \mathbf{I}^* \otimes \boldsymbol{\Sigma})$, *where* $\otimes$ *denotes the Kronecker product operator and* $\mathbf{I}^* = \mathrm{diag}(1/n_1, \cdots, 1/n_S)$.

Proof. Given $(n_1, \cdots, n_S)^\top$, $\overline{\mathbf{y}}_1, \cdots, \overline{\mathbf{y}}_S$ are independently distributed as $\mathcal{N}_C(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}/n_1), \cdots, \mathcal{N}_C(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}/n_S)$, and the result follows.

□

Property 2.3 $\widehat{\boldsymbol{\Sigma}}_0 = \mathbf{S}_{pooled} + \sum_{s=1}^{S} n_s \mathbf{d}_{0s}\mathbf{d}_{0s}^\top/N$, *where* $\mathbf{d}_{0s} = \overline{\mathbf{y}}_s - \boldsymbol{\mu}_{0s}$, $s = 1, \cdots, S$. *Also,* $\mathrm{E}(\widehat{\boldsymbol{\Sigma}}_0) = \boldsymbol{\Sigma}$.

Proof. The proof follows in a straightforward manner.

□

Property 2.4 *If $n_s > 0\ \forall s$, then $N\mathbf{S}_{pooled} \sim \mathcal{W}_C(\boldsymbol{\Sigma}, N-S)$ independently of $(n_1, \cdots, n_S)^\top$, where $\mathcal{W}_C(\boldsymbol{\Sigma}, N-S)$ is the Wishart distribution with scale matrix $\boldsymbol{\Sigma}$ and $N-S$ degrees of freedom.*

Proof. From standard results in normal distribution theory, $n_1\mathbf{S}_1, \cdots, n_S\mathbf{S}_S$ are independent $\mathcal{W}_C(\boldsymbol{\Sigma}, n_1-1), \cdots, \mathcal{W}_C(\boldsymbol{\Sigma}, n_S-1)$, conditional on $(n_1, \cdots, n_S)^\top$ such that $n_s > 0\ \forall s$. Using Theorem 3.4.3 of Mardia *et al.* (1979, p. 67), $N\mathbf{S}_{pooled}$, given $(n_1, \cdots, n_S)^\top$, has the given distribution. Since this is independent of $(n_1, \cdots, n_S)^\top$, the result follows.

□

Restrictions may be imposed on the model to reduce the number of unknown parameters in cases where S is far greater than N. One way to do this is to impose a log-linear structure for $\boldsymbol{\pi}$ and a linear model for $\boldsymbol{\mu}$ (Schafer, 1997).

2.3 One-Sample Location Hypotheses in the Mixed Data Case

Let $\boldsymbol{\theta}^\top = (\boldsymbol{\pi}^\top, \boldsymbol{\mu}^\top)$ be the vector of *location parameters* for $GLM(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The one-sample location problem

$$H : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad \text{against} \qquad K : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \tag{2.5}$$

is considered, with $\boldsymbol{\theta}_0^\top = (\boldsymbol{\pi}_0^\top, \boldsymbol{\mu}_0^\top)$ completely specified.

The likelihood ratio test (LRT) statistic as defined below (Mardia *et al.*, 1979, pp. 123-124) is adopted in the subsequent sections to construct statistical tests of (2.5) under various situations.

Definition 2.2 *Let the parameter space be* $\Theta = \Theta_H \cup \Theta_K$. *The likelihood ratio statistic* λ *for testing* $H : \boldsymbol{\theta} \in \Theta_H$ *against* $K : \boldsymbol{\theta} \in \Theta_K$ *is defined as*

$$\begin{aligned} \lambda &= \frac{\max_{\Theta_H} \mathcal{L}}{\max_{\Theta_K} \mathcal{L}} \\ &\equiv \frac{\widehat{\mathcal{L}}_H}{\widehat{\mathcal{L}}_K}. \end{aligned} \tag{2.6}$$

Equivalently, one may use the statistic $\log \lambda = \log \widehat{\mathcal{L}}_H - \log \widehat{\mathcal{L}}_K$.

2.3.1 Case of Known Covariance Matrix

Consider a general location model with location parameter $\boldsymbol{\theta}^\top = (\boldsymbol{\pi}^\top, \boldsymbol{\mu}^\top)$ and known covariance matrix $\boldsymbol{\Sigma}$.

Theorem 2.1 *Consider the hypotheses in (2.5).*

(*i*) *The* LRT *is of the form:*

$$\text{Reject } H \text{ iff } \sum_{s=1}^{S} \zeta_s > c,$$

for some critical value c, *where* $\zeta_s = (-2N/S) \log N - 2n_s \log(\pi_{0s}/n_s) + n_s \mathbf{d}_{0s}^\top \boldsymbol{\Sigma}^{-1} \mathbf{d}_{0s}$.

(*ii*) *For a level* α *test, the critical value* c_α *is obtained from*

$$\begin{aligned} \alpha &= \Pr\left(\sum_{s=1}^{S} \zeta_s > c_\alpha \mid \boldsymbol{\theta}_0 \right) \\ &= \sum_{n_1, \cdots, n_S} p(n_1, \cdots, n_S \mid \boldsymbol{\pi}_0) \Pr\left[\chi^2_{d_1} > c(n_1, \cdots, n_S) \right], \end{aligned} \tag{2.7}$$

where

$$p(n_1, \cdots, n_S \mid \boldsymbol{\pi}_0) = \frac{\prod_{s=1}^{S} \pi_{0s}^{n_s} / \prod_{s=1}^{S} n_s!}{\sum_{n_1,\cdots,n_S} \left(\prod_{s=1}^{S} \pi_{0s}^{n_s} / \prod_{s=1}^{S} n_s! \right)}, \tag{2.8}$$

with the summations taken over all $\{n_1, \cdots, n_S\}$ *such that* $n_s > 0 \ \forall s$ *and* $\sum_{s=1}^{S} n_s = N$, $\chi^2_{d_1}$ *is a* χ^2 *random variable with* $d_1 = CS$ *degrees of freedom, and* $c(n_1, \cdots, n_S) = c_\alpha / [-2N \log N - 2 \sum_{s=1}^{S} n_s \log(\pi_{0s}/n_s)]$.

(*iii*) *At any* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, *the power of the LRT is*

$$\Pr\left(\sum_{s=1}^{S} \zeta_s > c_\alpha \mid \boldsymbol{\theta} \right) = \sum_{n_1,\cdots,n_S} p(n_1, \cdots, n_S \mid \boldsymbol{\pi}) \times \Pr\left[\chi^2_{d_1, \Delta(n_1,\cdots,n_S)} > c(n_1, \cdots, n_S) \right], \tag{2.9}$$

where the summation is the same as in (*ii*), $p(n_1, \cdots, n_S \mid \boldsymbol{\pi})$ *is (2.8) with* $\boldsymbol{\pi}_0 = \boldsymbol{\pi}$, *and* $\chi^2_{d_1, \Delta(n_1,\cdots,n_S)}$ *is a noncentral* χ^2 *random variable with* $d_1 = CS$ *degrees of freedom and noncentrality parameter* $\Delta(n_1, \cdots, n_S) = \sum_{s=1}^{S} n_s (\boldsymbol{\mu}_s - \boldsymbol{\mu}_{0s})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_s - \boldsymbol{\mu}_{0s})$.

Proof. When $\boldsymbol{\Sigma}$ is known, the maximized log-likelihood under H is

$$\log \hat{\mathcal{L}}_H = \sum_{s=1}^{S} n_s \log \pi_{0s} - \frac{N}{2} \log \mid 2\pi\boldsymbol{\Sigma} \mid - \sum_{s=1}^{S} \frac{n_s}{2} \mathrm{tr} \left[\boldsymbol{\Sigma}^{-1} \left(\mathbf{S}_s + \mathbf{d}_{0s} \mathbf{d}_{0s}^\top \right) \right],$$

where $\mathbf{d}_{0s} = \overline{\mathbf{y}}_s - \boldsymbol{\mu}_{0s}$. Also, since K places no constraints on $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ is as given in (2.4) and

$$\log \hat{\mathcal{L}}_K = \sum_{s=1}^{S} n_s \log \left(\frac{n_s}{N} \right) - \frac{N}{2} \log \mid 2\pi\boldsymbol{\Sigma} \mid - \sum_{s=1}^{S} \frac{n_s}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}^{-1} \mathbf{S}_s \right).$$

Therefore, using (2.6), H is rejected if and only if $-2 \log \lambda = \sum_{s=1}^{S} \zeta_s > c$, for some c, which proves (*i*).

Noting that $n_1\mathbf{d}_{01}^\top\boldsymbol{\Sigma}^{-1}\mathbf{d}_{01},\cdots,n_S\mathbf{d}_{0S}^\top\boldsymbol{\Sigma}^{-1}\mathbf{d}_{0S}$ are independent and identically distributed χ^2 random variables each with C degrees of freedom, (2.7) in (*ii*) is obtained by using Theorem 2.5.2 of Mardia *et al.* (1979, p. 39) and the fact that $(n_1,\cdots,n_S)^\top$ has a truncated multinomial distribution with parameters N and $\boldsymbol{\pi}_0$ under H, subject to the condition $0 < n_s < N\ \forall s$, as given in (2.8).

Part (*iii*) is proved by noting that $n_1\mathbf{d}_{01}^\top\boldsymbol{\Sigma}^{-1}\mathbf{d}_{01},\cdots,n_S\mathbf{d}_{0S}^\top\boldsymbol{\Sigma}^{-1}\mathbf{d}_{0S}$ are independently distributed noncentral χ^2 random variables each with C degrees of freedom and respective noncentrality parameters $n_1(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{01})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{01}),\cdots,n_S(\boldsymbol{\mu}_S-\boldsymbol{\mu}_{0S})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_S-\boldsymbol{\mu}_{0S})$, and that for some given $\boldsymbol{\pi}$, $(n_1,\cdots,n_S)^\top$ has a truncated multinomial distribution with parameters N and $\boldsymbol{\pi}$, as shown in (2.9).

□

Remark 2.3.1 *The condition $n_s > 0\ \forall s$ (i.e., each state has at least 1 observation) is necessary so that all unknown parameters will be estimable. This results in a truncated multinomial distribution for $(n_1,\cdots,n_S)^\top$ given by (2.8) under H.*

Remark 2.3.2 *For the case where the states have different but known covariance matrices $\boldsymbol{\Sigma}_1,\cdots,\boldsymbol{\Sigma}_S$, Theorem 2.1 still applies except that ζ_s becomes* $\zeta_s = (-2N/S)\log N - 2n_s\log(\pi_{0s}/n_s) + n_s\mathbf{d}_{0s}^\top\boldsymbol{\Sigma}_s^{-1}\mathbf{d}_{0s}$.

Remark 2.3.3 *For a given level α, the critical value c_α in (2.7) can be computed quite easily using conventional methods (see, e.g., Faires and Burden, 1998).*

Remark 2.3.4 *If $S = 1$, the* LRT *statistic in Theorem 2.1 becomes $N(\overline{\mathbf{y}} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\overline{\mathbf{y}} - \boldsymbol{\mu}_0)$, i.e., the* LRT *in Theorem 2.1 reduces to the one-sample* LRT *for the multivariate normal distribution with known covariance matrix (Mardia* et al., *1979, p. 124). Thus, Theorem 2.1 generalizes the latter to the case of mixed binary and continuous data.*

The following corollary is obtained for the case $C = 1$ and $S = 2$ by applying Theorem 2.1 to the model considered in Example 2.1.

Corollary 2.1.1 *Consider the model in Example 2.1. Suppose σ^2 is known. The* LRT *of $H : (p, \mu_1, \mu_2)^\top = (p_0, \mu_{01}, \mu_{02})^\top$ against $K : (p, \mu_1, \mu_2)^\top \neq (p_0, \mu_{01}, \mu_{02})^\top$ in this case is of the form:*

$$\text{Reject } H \text{ iff } -2N\log N - 2n\log\left(\frac{p_0}{n}\right) - 2(N-n)\log\left(\frac{q_0}{N-n}\right) + \left[n(\overline{Y}_1 - \mu_{01})^2 + (N-n)(\overline{Y}_2 - \mu_{02})^2\right]/\sigma^2 > c_\alpha, \quad (2.10)$$

for some α-critical value c_α obtained from

$$\alpha = \frac{1}{1 - p_0^N - q_0^N} \sum_{n=1}^{N-1} \binom{N}{n} p_0^n q_0^{N-n} \Pr\left[\chi_2^2 > c(n)\right], \quad (2.11)$$

where $c(n) = c_\alpha/[-2N\log N - 2n\log(p_0/n) - 2(N-n)\log\{q_0/(N-n)\}]$.

The power of the test at $(p, \mu_1, \mu_2)^\top \neq (p, \mu_{01}, \mu_{02})^\top$ is

$$\frac{1}{1 - p^N - q^N} \sum_{n=1}^{N-1} \binom{N}{n} p^n q^{N-n} \Pr\left[\chi^2_{2,\Delta(n)} > c(n)\right], \quad (2.12)$$

where $\Delta(n) = [n(\mu_1 - \mu_{01})^2 + (N-n)(\mu_2 - \mu_{02})^2]/\sigma^2$.

Proof. The proof follows immediately from Theorem 2.1 by taking $C = 1$ and $S = 2$.

□

2.3.2 Case of Unknown Covariance Matrix

Consider a general location model with location parameter $\boldsymbol{\theta}^\top = (\boldsymbol{\pi}^\top, \boldsymbol{\mu}^\top)$ and unknown covariance matrix $\boldsymbol{\Sigma}$. This is usually the case in many applied studies where knowledge about $\boldsymbol{\Sigma}$ is not available.

Theorem 2.2 *Consider the hypotheses in (2.5).*

(*i*) *The* LRT *is of the form:*

$$\text{Reject } H \text{ iff } \sum_{s=1}^{S} \zeta_s > c,$$

for some critical value c, where

$$\zeta_s = a(n_1, \cdots, n_S; \boldsymbol{\pi}_0) \left[\frac{1}{S} + \frac{n_s}{N} \mathbf{d}_{0s}^\top \mathbf{S}_{pooled}^{-1} \mathbf{d}_{0s} \right]$$

and $a(n_1, \cdots, n_S; \boldsymbol{\pi}_0) = N^{-2} \prod_{s=1}^{S} (n_s/\pi_{0s})^{2n_s/N}$.

(*ii*) *For a level* α *test, the critical value* c_α *is obtained from*

$$\begin{aligned} \alpha &= \Pr\left(\sum_{s=1}^{S} \zeta_s > c_\alpha \mid \boldsymbol{\theta}_0 \right) \\ &= \sum_{n_1, \cdots, n_S} p(n_1, \cdots, n_S \mid \boldsymbol{\pi}_0) \Pr\left[U^{(M)} > c(n_1, \cdots, n_S) \right], \end{aligned} \tag{2.13}$$

where the summation is taken over all $\{n_1, \cdots, n_S\}$ *such that* $n_s > 0 \; \forall s$ *and* $\sum_{s=1}^{S} n_s = N$, $p(n_1, \cdots, n_S \mid \boldsymbol{\pi}_0)$ *is defined in (2.8),* $c(n_1, \cdots, n_S) = a(n_1, \cdots, n_S; \boldsymbol{\pi}_0) c_\alpha - 1$, *and* $U^{(M)}$ *has the same distribution as that, under* H *and conditional on* $(n_1, \cdots, n_S)^\top$, *of the sum of non-zero roots* $\xi_1, \cdots, \xi_M$ *of the following determinantal equation:*

$$\left| \sum_{s=1}^{S} \frac{n_s}{N} \mathbf{d}_{0s} \mathbf{d}_{0s}^\top - \xi \mathbf{S}_{pooled} \right| = 0.$$

Proof. From § 2.2.2, the MLEs of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are, respectively, $\boldsymbol{\mu}_0$ and $\widehat{\boldsymbol{\Sigma}}_0$ under H, and $\overline{\mathbf{y}}$ and $\mathbf{S}_{pooled}$ under K. Now from (2.1), it can be deduced that

$$\begin{aligned}
\mathcal{L}_H &= \prod_{s=1}^{S} \pi_{0s}^{n_s} \mid 2\pi\widehat{\boldsymbol{\Sigma}}_0 \mid^{-N/2} \exp\left(-\frac{CN}{2}\right), \\
\mathcal{L}_K &= \prod_{s=1}^{S} \left(\frac{n_s}{N}\right)^{n_s} \mid 2\pi\mathbf{S}_{pooled} \mid^{-N/2} \exp\left(-\frac{CN}{2}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\lambda^{-2/N} &= a(n_1, \cdots, n_S; \boldsymbol{\pi}_0) \times \frac{\mid \widehat{\boldsymbol{\Sigma}}_0 \mid}{\mid \mathbf{S}_{pooled} \mid} \\
&= \sum_{s=1}^{S} \zeta_s, \qquad (2.14)
\end{aligned}$$

using Property 2.3 and A2.3m of Mardia *et al.* (1979, p. 458), and H is rejected for large values of (2.14). This proves (i).

Now consider (ii), and assume H holds and $(n_1, \cdots, n_S)^\top$ is fixed. It follows that $\sqrt{n_s}\mathbf{d}_{0s} \sim \mathcal{N}_C(\mathbf{0}, \boldsymbol{\Sigma})$ (Property 2.2), so that $\mathbf{M}_1 = \sum_{s=1}^{S} n_s \mathbf{d}_{0s}\mathbf{d}_{0s}^\top \sim \mathcal{W}_C(\boldsymbol{\Sigma}, S)$, independently of $\mathbf{M}_2 = N\mathbf{S}_{pooled} \sim \mathcal{W}_C(\boldsymbol{\Sigma}, N - S)$ (Property 2.4). Write (2.14) as

$$\begin{aligned}
\lambda^{-2/N} &= a(n_1, \cdots, n_S; \boldsymbol{\pi}_0) \left[1 + \operatorname{tr}\left(\mathbf{M}_1\mathbf{M}_2^{-1}\right)\right] \\
&= a(n_1, \cdots, n_S; \boldsymbol{\pi}_0) \left(1 + U^{(M)}\right).
\end{aligned}$$

Using the fact that the trace of a matrix is equal to the sum of its eigenvalues (Mardia *et al.*, 1979, p. 467), it is clear that $U^{(M)}$ has the same distribution as $\sum_{m=1}^{M} \xi_m$, where $\xi_1 \neq 0, \cdots, \xi_M \neq 0$ satisfy

$$\begin{aligned}
0 &= \left|\mathbf{M}_1\mathbf{M}_2^{-1} - \xi\mathbf{I}_C\right| \\
&= \left|\sum_{s=1}^{S} \frac{n_s}{N}\mathbf{d}_{0s}\mathbf{d}_{0s}^\top - \xi\mathbf{S}_{pooled}\right|.
\end{aligned}$$

Finally, the expression in (2.13) is obtained by noting that $(n_1, \cdots, n_S)^\top$ has a truncated multinomial distribution with parameters N and $\boldsymbol{\pi}_0$ under H, subject to the condition that $0 < n_s < N \ \forall s$.

□

Remark 2.3.5 *Remarks 2.3.1 and 2.3.3 for Theorem 2.1 also apply to this case.*

Remark 2.3.6 *It is assumed that the Wishart distribution is nonsingular, i.e., $N \geq S + C$, so that $\mathbf{S}_{\text{pooled}}^{-1}$ exists with probability 1 (Dykstra, 1970). This holds if $\forall s$, $n_s \geq C$ so that $n_s\mathbf{S}_s$ has a nonsingular Wishart distribution, $s = 1, \cdots, S$. In this case, $M = \text{minimum}(S, N - S)$.*

Remark 2.3.7 *The statistic $U^{(M)}$ is known in the literature as Hotelling's generalized T^2 statistic (Hotelling, 1951). Its null, or central, distribution was derived by Hotelling (1951) for $C = 2$ and by Krishnaiah and Chang (1972), Pillai and Young (1971), and Davis (1970) for $C > 2$. Mijares (1990), McKeon (1974), Hughes and Saw (1972), and Pillai and Sampson (1959) provide various approximations to this distribution. A noncentral distribution arises under K, since $\sum_{s=1}^{S} n_s\mathbf{d}_{0s}\mathbf{d}_{0s}^\top$ has a noncentral Wishart distribution in this case.*

Remark 2.3.8 *If $S = 1$, the* LRT *statistic in Theorem 2.2 becomes $(1 + \mathbf{d}_0^\top\mathbf{S}_{\text{pooled}}^{-1}\mathbf{d}_0)$, i.e., the* LRT *in Theorem 2.2 reduces to the one-sample LRT for a multivariate normal distribution with unknown $\boldsymbol{\Sigma}$ (Mardia* et al., *1979, p. 125).*

The following corollary for the case $C = 1$ and $S = 2$ is a special case of Theorem 2.2 as applied to the model considered in Example 2.1.

Corollary 2.2.1 *Consider the model in Example 2.1. Suppose σ^2 is unknown. The* LRT *of $H : (p, \mu_1, \mu_2)^\top = (p_0, \mu_{01}, \mu_{02})^\top$ against $K : (p, \mu_1, \mu_2)^\top \neq (p_0, \mu_{01}, \mu_{02})^\top$ in this case is of the form:*

$$\text{Reject } H \text{ iff } a(n; p_0)\left[1 + \frac{n(\overline{Y}_1 - \mu_{01})^2 + (N-n)(\overline{Y}_2 - \mu_{02})^2}{Q(\overline{Y}_1, \overline{Y}_2)}\right] > c_\alpha, \quad (2.15)$$

for some α-critical value c_α obtained from

$$\alpha = \frac{1}{1 - p_0^N - q_0^N} \sum_{n=1}^{N-1} \binom{N}{n} p_0^n q_0^{N-n} \Pr\left[F_{2,N-2} > c(n)\right], \quad (2.16)$$

where $F_{2,N-2}$ is an F random variable with $(2, N-2)$ degrees of freedom, $a(n; p_0) = (1/N^2)(n/p_0)^{2n/N}[(N-n)/q_0]^{2-2n/N}$ and $c(n) = (N-2)[c_\alpha/a(n; p_0) - 1]/2$.

The power of the test at $(p, \mu_1, \mu_2)^\top \neq (p_0, \mu_{01}, \mu_{02})^\top$ is given by

$$\frac{1}{1 - p^N - q^N} \sum_{n=1}^{N-1} \binom{N}{n} p^n q^{N-n} \Pr\left[F_{2,N-2}^{\Delta(n)} > c(n)\right], \quad (2.17)$$

where $F_{2,N-2}^{\Delta(n)}$ is a noncentral $F_{2,N-2}$ random variable with noncentrality parameter $\Delta(n) = [n(\mu_1 - \mu_{01})^2 + (N-n)(\mu_2 - \mu_{02})^2]/\sigma^2$.

Proof. The test statistic in (2.15) is easily obtained from Theorem 2.2 by taking $S = 2$ and $C = 1$.

The null distribution in (2.16) is derived by noting that, under H and conditional on n, $n(\overline{Y}_1 - \mu_{01})^2/\sigma^2$, $(N-n)(\overline{Y}_2 - \mu_{02})^2/\sigma^2$, and $Q(\overline{Y}_1, \overline{Y}_2)/\sigma^2$

are independent χ^2_1, χ^2_1 and χ^2_{N-2} random variables, respectively. Hence,

$$\frac{N-2}{2} \times \frac{n(\overline{Y}_1 - \mu_{01})^2 + (N-n)(\overline{Y}_2 - \mu_{02})^2}{Q(\overline{Y}_1, \overline{Y}_2)} = \frac{\chi^2_2/2}{\chi^2_{N-2}/(N-2)}$$
$$\sim F_{2,N-2},$$

under H and conditional on n. The rest of (2.16) follows from the fact that n has a binomial distribution with parameters (N, p_0) under H, truncated at 0 and N.

Expression (2.17) follows from the non-null distribution, which is obtained in a similar fashion, except that the F variable now becomes a noncentral F variable with the same degrees of freedom and noncentrality parameter $\Delta(n) = [n(\mu_1 - \mu_{01})^2 + (N-n)(\mu_2 - \mu_{02})^2]/\sigma^2$ (Johnson and Kotz, 1970, pp. 189-190).

□

Critical values for the LRT in Corollary 2.2.1 were computed in S-PLUS and are displayed in Table 2.1 (de Leon and Carrière, 2000) for various values of N and p_0 at levels $\alpha = 0.01, 0.05$. Note that Table 2.1 may be used as well for $1 - p_0 = 0.75, 0.9$, and 0.95, as the critical values c_α are the same for these cases. For example, the critical value at $\alpha = 0.01$ when $p_0 = 0.05$ with $N = 30$ (i.e., $c_\alpha = 1.453$) is exactly the same as that when $p_0 = 0.95$.

Figure 2.1 displays several plots of the power function in (2.17) for various fixed true values μ_1, μ_2 and null values μ_{01}, μ_{02} of the state means, with $p_0 = 0.5$. Plots $(a) - (b)$ have $N = 25$ while $(c) - (d)$ have $N = 50$. It is clear from

Figure 2.1 that the power of the LRT increases with N as well as with the distance between the null and true values of the state means.

Similarly, contour plots of (2.17) for a range of values of (μ_1, μ_2) are shown in Figure 2.2, with $p = 0.5, \sigma^2 = 1$ and $N = 25$. The null values considered are $\mu_{01} = 0, \mu_{02} = 0.5$ and $p_0 = 0.25, 0.5$. The contour levels are generally high (especially for the top plot in Figure 2.2) around, but decreasing as they approach, the point $(\mu_1, \mu_2) = (0, 0.5)$. This indicates a funnel-like shape for the power surface, with saddle point at $(\mu_1, \mu_2) = (0, 0.5)$.

2.3.3 Properties of the Likelihood Ratio Test

In what follows, two optimal properties, namely, *consistency* and *unbiasedness*, are proved for the LRTs derived in § 2.3.1-2.3.2.

Property 2.5 *The* LRT *in Theorem 2.1 is consistent. The same holds for that in Theorem 2.2, provided* $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.

Proof. For Theorem 2.1, note that $-2N \log N - 2\sum_{s=1}^{S} n_s \log(\pi_{0s}/n_s) \to 0$ almost surely as $N \to \infty$. Therefore it follows that $c(n_1, \cdots, n_S) \to \infty$, and consistency follows.

For Theorem 2.2, c_α satisfies $\Pr\left(\sum_{s=1}^{S} \zeta_s > c_\alpha \mid \boldsymbol{\theta}_0\right) = \alpha$. Because $\chi^2_{d_2}/d_2 \to 1$ and $a(n_1, \cdots, n_S; \boldsymbol{\pi}_0) \to 1$ almost surely, it follows that $d_2(c_\alpha - 1) \to c_0$ such that $\Pr(\chi^2_{d_1} \leq c_0) = 1-\alpha$, where $d_1 = CS$ and $d_2 = N-S-C+1$. Thus, $c_\alpha \to 1$ as $N \to \infty$. By the strong law of large numbers,

$$\frac{d_2}{N d_1} \sum_{s=1}^{S} n_s \mathbf{d}_{0s}^\top \mathbf{S}_{pooled}^{-1} \mathbf{d}_{0s} \to \sum_{s=1}^{S} (\boldsymbol{\mu}_s - \boldsymbol{\mu}_{0s})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_s - \boldsymbol{\mu}_{0s}),$$

almost surely. For $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, $\sum_{s=1}^{S}(\boldsymbol{\mu}_s - \boldsymbol{\mu}_{0s})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_s - \boldsymbol{\mu}_{0s}) > 0$ and the LRT is consistent.

□

Property 2.6 *The* LRT*s in Theorems 2.1 and 2.2 are both unbiased.*

Proof. Because the distribution functions of $\chi^2_{d_1,\Delta(n_1,\cdots,n_s)}$ and $F^{\Delta(n_1,\cdots,n_S)}_{d_1,d_2}$ are decreasing in $\Delta(n_1,\cdots,n_S)\ \forall n_1,\cdots,n_S$ (Johnson and Kotz, 1970, pp. 135,193), it follows that $\Pr[\chi^2_{d_1,\Delta(n_1,\cdots,n_S)} \leq c] \leq \Pr[\chi^2_{d_1} \leq c]$, and $\Pr[F^{\Delta(n_1,\cdots,n_S)}_{d_1,d_2} \leq c] \leq \Pr[F_{d_1,d_2} \leq c]$, for any constant c. Therefore,

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\pi}}\{\Pr[\chi^2_{d_1,\Delta(n_1,\cdots,n_S)} > c(n_1,\cdots,n_S)]\} &\geq \mathrm{E}_{\boldsymbol{\pi}}\{\Pr[\chi^2_{d_1} > c(n_1,\cdots,n_S)]\},\\
\mathrm{E}_{\boldsymbol{\pi}}\{\Pr[F^{\Delta(n_1,\cdots,n_S)}_{d_1,d_2} > c(n_1,\cdots,n_S)]\} &\geq \mathrm{E}_{\boldsymbol{\pi}}\{\Pr[F_{d_1,d_2} > c(n_1,\cdots,n_S)]\},
\end{aligned}$$

where the expectations are taken with respect to the truncated multinomial distributions with parameters $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_0$. This implies that the power achieves its minimum at $\boldsymbol{\theta}_0$, and the LRTs are unbiased (Anderson, 1984, p. 362).

□

The above proofs include as special cases those for the case $C = 1$ and $S = 2$, details of which are found in de Leon and Carrière (2000).

2.3.4 Simulation Results

In this section, the power of the LRT derived in Corollary 2.2.1 is investigated and its performance compared against that of the *separate test* approach, which treats the binary and continuous variables individually and tests their

parameters separately. The relative power superiority of the LRT developed in this chapter compared to that of the separate test may be anticipated as the former utilizes the information about the dependency between the variables $\mathbf{x} = (X_1, X_2)^\top$ and Y in the general location model considered in Example 2.1. The actual power values of LRT are presented here to confirm this conjecture as well as to show the relative merits of LRT over the separate test.

The separate test approach entails carrying out simultaneous tests of the following hypotheses:

$$\begin{aligned} H_0 : p = p_0 \quad &\text{against} \quad K_0 : p \neq p_0, \\ H_1 : \mu_1 = \mu_{01} \quad &\text{against} \quad K_1 : \mu_1 \neq \mu_{01}, \\ H_2 : \mu_2 = \mu_{02} \quad &\text{against} \quad K_2 : \mu_2 \neq \mu_{02}. \end{aligned} \tag{2.18}$$

Note that $H = \bigcap_{j=0}^{2} H_j$. The first pair above is tested using the exact binomial test while the latter two are tested using the standard one-sample t-test, using the pooled sample variance to estimate σ^2. If at least one null hypothesis in (2.18) is rejected, then the null hypothesis H in (2.5) is rejected. To control the overall level of the tests, a *Bonferroni adjustment* (Pocock *et al.*, 1987) of the level of each test is made by dividing the nominal level α by 3.

Because the power function for the separate test is not known analytically, power values are directly calculated only for LRT using the power function given in (2.17), and Monte Carlo simulation is used for the separate test. In the first simulation experiment, samples of moderate sizes $N = 15$ and 25 were generated from the general location models with scale parameter $\sigma^2 = 25$ and the following values for the location parameter $\boldsymbol{\theta} = (p, \mu_1, \mu_2)^\top$: (a)

$(0.35, 50, 25)^\top$, (b) $(0.35, 52.5, 22.5)^\top$, (c) $(0.4, 55, 22.5)^\top$, and (d) $(0.4, 55, 20)^\top$. In each case, the null parameter $\boldsymbol{\theta}_0$ was taken to be $(0.3, 50, 25)^\top$. To maximize the advantage of using LRT, the difference in the two mean values should be quite large and this influenced the choice of the parameters above. From Olkin and Tate (1961) and Tate (1954), the correlation (also referred to as *point-biserial correlation*) between $\mathbf{x}$ and Y is given by

$$\rho_{\mathbf{x},Y} = (\mu_1 - \mu_2)\sqrt{\frac{pq}{\sigma^2 + pq(\mu_1 - \mu_2)^2}}.$$

Therefore, under the location model, $\mathbf{x}$ and Y become uncorrelated (in fact, independent) if the two state means are equal. Conversely, the dependency becomes stronger when they are far from each other, and it is precisely where LRT is expected to outperform the separate test. The performance of LRT will be equivalent to that of the separate test as the dependency between $\mathbf{x}$ and Y becomes negligible. Note that the values of $\rho_{\mathbf{x},Y}$ for the general location models considered are generally high, ranging from 0.922 to 0.96 in cases (a)-(d). It is equal to 0.916 in case (0).

In the second simulation experiment, the null model is taken as the general location model with $\boldsymbol{\theta}_0 = (0.5, 50, 45)^\top$ and $\sigma^2 = 25$. For power comparisons, samples of the same sizes $N = 15$ and 25 were generated from the general location models with $\sigma^2 = 25$ and location parameters $\boldsymbol{\theta}$ equal to (a)' $(0.45, 50, 42.5)^\top$, (b)' $(0.45, 52.5, 42.5)^\top$, (c)' $(0.55, 52.5, 42.5)^\top$, and (d)' $(0.55, 55, 47.5)^\top$. Note that unlike in the first experiment where the values of $\rho_{\mathbf{x},Y}$ were generally very high, those for the second experiment range from a low of 0.447 (for the null model) to a high of 0.705.

Tables 2.2 and 2.3 present the powers and the empirical sizes. All samples were generated using S-PLUS, with 10,000 repetitions in each case. The entries in Cases (0) and (0)' correspond to the situation when the null hypothesis is true, and hence give the levels (empirical in the case of the separate test) of the tests. They indicate that the actual level of LRT is exactly at the nominal level, while that for the separate test tended to be slightly conservative.

It is clear from Tables 2.2 and 2.3 that the performance of LRT is superior to that of the separate test, as the power values are generally much higher for the former compared with those of the latter. This is true even in the case of a slight departure from the true value as in (a) and (b) in the first experiment, and (a)' in the second. The comparison is most favorable to LRT when $\alpha = 0.01$, and especially when $N = 15$. This can be explained mainly by the fact that the separate test is a conservative method and becomes especially so for small sample sizes and high correlation between $\mathbf{x}$ and Y. The findings reported here are in general agreement with those reported by Pocock *et al.* (1987) for continuous variables.

2.4 Extension to G-Sample Case

The multi-sample location problem

$$H : \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_G \qquad \text{against} \qquad K : \textit{at least 1 inequality}. \tag{2.19}$$

is the focus of this section. The interest is on constructing statistical tests of (2.19) based on $G \geq 2$ independent random samples from $GLM(\boldsymbol{\pi}_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, $g = 1, \cdots, G$. Note that (2.19) is the *one-way multivariate analysis of variance*

(MANOVA) problem involving mixed binary and continuous populations. For convenience, denote by $(\mathbf{x}_{gsi}^\top, \mathbf{y}_{gsi}^\top)^\top$ the ith observation belonging to state s in population g, and put $\boldsymbol{\theta}_g^\top = (\boldsymbol{\pi}_g^\top, \boldsymbol{\mu}_g^\top)$.

Previous works on mixed MANOVA problems include those of Afifi and Elashoff (1969), Pocock *et al.* (1987), and more recently, that of Morales *et al.* (1998). Pocock *et al.* (1987) obtained global tests for comparing treatment effects in clinical trials. They extended O'Brien's (1984) work to binary and survival data and established the asymptotic normality of the test statistics. Morales *et al.* (1998) proposed a class of dissimilarity measures among several independent populations each described by the general location model, and applied it to hypothesis-testing problems similar to (2.19).

The approach adopted here is most similar to that of Afifi and Elashoff (1969), who considered the two-sample (i.e., $G = 2$) location problem and derived an information-theoretic test, in addition to the LRT, for the problem. Here, as in § 2.3 and Afifi and Elashoff (1969), the likelihood ratio approach is adopted to derive global statistical tests of (2.19). The case of complete *homogeneity* of the populations where $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_G$, is discussed first. This can be viewed as a generalization of the classical MANOVA problem to the mixed-data setting.

As well, two scenarios where *heterogeneity* can arise are also considered. In the first, it is assumed that the covariance matrix is the same across populations but varies across states. In the second, intra-population but not inter-population homogeneity is assumed. This latter case can be viewed as a more

general case of the *Behrens-Fisher problem* (Mardia *et al.*, 1979, p. 142-144) involving mixed data.

2.4.1 Case of Complete Homogeneity

Given G independent random samples from the general location models with location parameters $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_G$ and common covariance matrix $\boldsymbol{\Sigma}$, the likelihood function can be written as

$$\begin{aligned} \mathcal{L} &= \prod_{g=1}^{G}\prod_{s=1}^{S} \pi_{gs}^{n_{gs}} \mid 2\pi\boldsymbol{\Sigma} \mid^{-N/2} \\ &\quad \times \exp\left\{-\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\sum_{g=1}^{G}\sum_{s=1}^{S}\frac{n_{gs}}{2}\left(\mathbf{S}_{gs}+\mathbf{d}_{gs}\mathbf{d}_{gs}^{\top}\right)\right]\right\}, \end{aligned} \tag{2.20}$$

where $\mathbf{d}_{gs} = \overline{\mathbf{y}}_{gs} - \boldsymbol{\mu}_{gs}$, $\mathbf{S}_{gs}$ and $\overline{\mathbf{y}}_{gs}$ are the sample mean and sample covariance matrix (uncorrected for bias), respectively, of the continuous observations belonging to state s in population g, $N = \sum_{g=1}^{G} n_{g\cdot} = \sum_{s=1}^{S} n_{\cdot s}$, $n_{g\cdot} = \sum_{s=1}^{S} n_{gs}$, and $n_{\cdot s} = \sum_{g=1}^{G} n_{gs}$, with n_{gs} the number of observations belonging to state s in population g. Note that (2.20) corresponds to $\mathcal{L}_K$, the likelihood under K in (2.19), where no constraints are imposed on $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_G$.

Derivation of the (unrestricted) MLEs of $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_G$ and $\boldsymbol{\Sigma}$ follows from standard results on multinomial (Bishop *et al.*, 1975) and multivariate normal distributions (Mardia *et al.*, 1979), and are given by

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_g^{\top} &= (\widehat{\boldsymbol{\pi}}_{g\cdot}^{\top}, \overline{\mathbf{y}}_{g\cdot}^{\top}) \quad (g = 1, \cdots, G) \\ N\widehat{\boldsymbol{\Sigma}} &= \sum_{g=1}^{G} n_{g\cdot}\mathbf{S}_{g\cdot}. \end{aligned} \tag{2.21}$$

For population g, $\widehat{\boldsymbol{\pi}}_{g\cdot}^{\top} = (n_{g1}/n_{g\cdot}, \cdots, n_{gS}/n_{g\cdot})$, $\overline{\mathbf{y}}_{g\cdot} = \sum_{s=1}^{S} n_{gs}\overline{\mathbf{y}}_{gs}/n_{g\cdot}$, and $n_{g\cdot}\mathbf{S}_{g\cdot} = \sum_{s=1}^{S} n_{gs}\mathbf{S}_{gs}$, similar to (2.4). The matrix $N\widehat{\boldsymbol{\Sigma}}$ is analogous to the

within-group sum of squares and product (SSP) matrix (Mardia *et al.*, 1979, p. 138) in MANOVA.

Under the hypothesis H of complete homogeneity, the G samples can be treated as constituting one sample from $GLM(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Therefore, the likelihood function $\mathcal{L}_H$ is exactly as given in (2.1) and the MLEs are then given by

$$\begin{aligned} \widehat{\widehat{\boldsymbol{\pi}}}^{\top} &= (n_{\cdot 1}/N, \cdots, n_{\cdot S}/N); \\ \\ \widehat{\widehat{\boldsymbol{\mu}}}^{\top} &= (\sum_{g=1}^{G} n_{g1}\overline{\mathbf{y}}_{g1}^{\top}/n_{\cdot 1}, \cdots, \sum_{g=1}^{G} n_{gS}\overline{\mathbf{y}}_{gS}^{\top}/n_{\cdot S}) \\ &= (\overline{\mathbf{y}}_{\cdot 1}^{\top}, \cdots, \overline{\mathbf{y}}_{\cdot S}^{\top}) \qquad (2.22) \\ &\equiv \overline{\mathbf{y}}^{\top}; \\ \\ N\widehat{\widehat{\boldsymbol{\Sigma}}} &= \sum_{g=1}^{G}\sum_{s=1}^{S}\sum_{i=1}^{n_{gs}} (\mathbf{y}_{gsi} - \overline{\mathbf{y}}_{\cdot s})(\mathbf{y}_{gsi} - \overline{\mathbf{y}}_{\cdot s})^{\top} \\ &\equiv N\mathbf{S}. \end{aligned}$$

The matrix $N\mathbf{S}$ is called the *total SSP* matrix (Mardia *et al.*, 1979, p. 138) in MANOVA.

Theorem 2.3 *Consider the hypotheses in (2.19).*

(*i*) *The* LRT *statistic is given by*

$$\lambda^{2/N} = b(n_{11}, \cdots, n_{GS}) \left| \mathbf{I}_C + \frac{\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{H}}{N} \right|^{-1}, \qquad (2.23)$$

where

$$b(n_{11}, \cdots, n_{GS}) = \left[\prod_{g=1}^{G}\prod_{s=1}^{S} \frac{n_{g\cdot}^{n_{g\cdot}} n_{\cdot s}^{n_{gs}}}{N^{n_{g\cdot}} n_{gs}^{n_{gs}}} \right]^{2/N}$$

and

$$\mathbf{H} = \sum_{g=1}^{G}\sum_{s=1}^{S} n_{gs}(\overline{\mathbf{y}}_{gs} - \overline{\mathbf{y}}_{\cdot s})(\overline{\mathbf{y}}_{gs} - \overline{\mathbf{y}}_{\cdot s})^{\top}.$$

The null hypothesis H is then rejected if and only if $\lambda^{2/N} < c$, for some critical value c.

(*ii*) *For a level α test, the critical value c_α is obtained from*

$$\begin{aligned} \alpha &= \Pr\left(\lambda^{2/N} < c_\alpha \mid \mathbf{t}, H\right) \\ &= \sum_{n_{11},\cdots,n_{GS}} p(n_{11},\cdots,n_{GS} \mid \mathbf{t}, H) \\ &\quad \times \Pr\left[\Lambda_{w_1,w_2,w_3} \leq c\left(n_{11},\cdots,n_{GS}\right)\right], \end{aligned} \tag{2.24}$$

where

$$\begin{aligned} p(n_{11},\cdots,n_{GS} \mid \mathbf{t}, H) &= \prod_{s=1}^{S} n_{\cdot s}! \prod_{g=1}^{G}\left(n_{g\cdot}!/\prod_{s'=1}^{S} n_{gs'}!\right) / \\ &\quad \sum_{n_{11},\cdots,n_{GS}} \prod_{s=1}^{S} n_{\cdot s}! \prod_{g=1}^{G}\left(n_{g\cdot}!/\prod_{s'=1}^{S} n_{gs'}!\right), \end{aligned} \tag{2.25}$$

with the summations taken over all $\{n_{11},\cdots,n_{GS}\}$ such that $n_{gs} > 0$ $\forall g, s$ and $\sum_{g=1}^{G} n_{gs} = n_{\cdot s}$ $\forall s$, Λ_{w_1,w_2,w_3} has the Wilks' λ-distribution with parameters $w_1 = C, w_2 = N - GS$, and $w_3 = S(G-1)$ provided $N \geq C + SG$, and $\mathbf{t}^{\top} = (n_{\cdot 1},\cdots,n_{\cdot S})$, $c(n_{11},\cdots,n_{GS}) = c_\alpha / b(n_{11},\cdots,n_{GS})$.

Proof. The LRT is easily derived from previous discussions. The MLEs under H are given in (2.22), since the data can be viewed under H as a single random sample whereas those under the alternative K are given in (2.21). Using (2.20),

the LRT statistic is thus

$$\begin{aligned}\lambda^{2/N} &= b(n_{11},\cdots,n_{GS}) \times \frac{|\widehat{\boldsymbol{\Sigma}}|}{|\mathbf{S}|} \\ &= b(n_{11},\cdots,n_{GS}) \times \left|\mathbf{I}_C + \frac{\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{H}}{N}\right|^{-1},\end{aligned}$$

since $N\mathbf{S} = N\widehat{\boldsymbol{\Sigma}} + \mathbf{H}$, where $\mathbf{H} = \sum_{g=1}^{G}\sum_{s=1}^{S} n_{gs}(\overline{\mathbf{y}}_{gs} - \overline{\mathbf{y}}_{\cdot s})(\overline{\mathbf{y}}_{gs} - \overline{\mathbf{y}}_{\cdot s})^{\top}$. The LRT thus rejects H for small values of $\lambda^{2/N}$. This proves (i).

Part (ii) is proved in two steps. First, the conditional distribution of $|\mathbf{I}_C + \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{H}/N|^{-1}$, given $\{n_{11},\cdots,n_{GS}\}$, is obtained. Following Mardia *et al.* (1979, pp. 138-139), let

$$\mathbf{W}_s = \begin{pmatrix} \mathbf{W}_{1s} \\ \vdots \\ \mathbf{W}_{Gs} \end{pmatrix},$$

for $s = 1,\cdots,S$, where $\mathbf{W}_{gs}$ represents the n_{gs} observations belonging to state s from population g, $g = 1,\cdots,G$. From Mardia *et al.* (1979, p. 139), it can be shown that

$$\sum_{g=1}^{G} n_{gs}\mathbf{S}_{gs} = \mathbf{W}_s^{\top}\mathbf{C}_1\mathbf{W}_s, \quad \sum_{g=1}^{G} n_{gs}(\overline{\mathbf{y}}_{gs} - \overline{\mathbf{y}}_{\cdot s})(\overline{\mathbf{y}}_{gs} - \overline{\mathbf{y}}_{\cdot s})^{\top} = \mathbf{W}_s^{\top}\mathbf{C}_2\mathbf{W}_s,$$

where $\mathbf{C}_1 = \sum_{g=1}^{G}[\mathrm{diag}(\mathbf{1}_g) - \mathbf{1}_g\mathbf{1}_g^{\top}/n_{gs}]$ and $\mathbf{C}_2 = \sum_{g=1}^{G}(\mathbf{1}_g\mathbf{1}_g^{\top}/n_{gs} - \mathbf{1}\mathbf{1}^{\top}/n_{\cdot s})$, with $\mathbf{1}_g$ denoting the $n_{\cdot s} \times 1$ vector with 1 in the positions corresponding to the gth sample and 0 elsewhere, and $\mathbf{1} = \sum_{g=1}^{G}\mathbf{1}_g$.

Now under H, $\mathbf{W}_s$ is a sample from $\mathcal{N}_C(\boldsymbol{\mu}_s, \boldsymbol{\Sigma})$ whence, by Theorems 3.4.4 and 3.4.5 of Mardia *et al.* (1979, pp. 68-70),

$$\begin{aligned}\mathbf{W}_s^{\top}\mathbf{C}_1\mathbf{W}_s &\sim \mathcal{W}_C(\boldsymbol{\Sigma}, n_{\cdot s} - G) \\ \mathbf{W}_s^{\top}\mathbf{C}_2\mathbf{W}_s &\sim \mathcal{W}_C(\boldsymbol{\Sigma}, G-1),\end{aligned}$$

and, moreover, $\mathbf{W}_s^\top \mathbf{C}_1 \mathbf{W}_s$ and $\mathbf{W}_s^\top \mathbf{C}_2 \mathbf{W}_s$ are independent. Since $\mathbf{W}_1, \cdots, \mathbf{W}_S$ are independent given $\{n_{11}, \cdots, n_{GS}\}$, it follows from Theorem 3.4.4 of Mardia *et al.* (1979, p. 67) that

$$\begin{aligned} N\widehat{\boldsymbol{\Sigma}} &\sim \mathcal{W}_C(\boldsymbol{\Sigma}, N - SG) \\ \mathbf{H} &\sim \mathcal{W}_C[\boldsymbol{\Sigma}, S(G-1)]. \end{aligned}$$

Given $\{n_{11}, \cdots, n_{GS}\}$, it follows from Definition 3.7.1 of Mardia *et al.* (1979, p. 81) that $|\mathbf{I}_C + \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{H}/N|^{-1} \sim \Lambda_{w_1,w_2,w_3}$ under H, where Λ_{w_1,w_2,w_3} has the Wilks' λ-distribution with parameters $w_1 = C, w_2 = N - GS$, and $w_3 = S(G-1)$, provided $N \geq C + GS$.

Next, the conditional joint distribution of $\mathbf{n}_1, \cdots, \mathbf{n}_G$ given $\mathbf{t} = \sum_{g=1}^G \mathbf{n}_g$ is derived under H, where $\mathbf{n}_g^\top = (n_{g1}, \cdots, n_{gS})$. To do this, observe that $\mathbf{n}_g$ is multinomial with parameters $n_{g\cdot}$ and $\boldsymbol{\pi}_g$ and, under H, $\mathbf{t}$ is also multinomial with parameters N and $\boldsymbol{\pi}_1 = \cdots = \boldsymbol{\pi}_G = \boldsymbol{\pi}$. Thus, under H,

$$p^*(\mathbf{n}_1, \cdots, \mathbf{n}_G \mid \mathbf{t}, H) = \frac{\prod_{g=1}^G \left(n_{g\cdot}! / \prod_{s=1}^S n_{gs}! \right)}{N! / \prod_{s=1}^S n_{\cdot s}!}. \tag{2.26}$$

The expression in (2.25) now follows, and the joint distribution of $\mathbf{n}_1, \cdots, \mathbf{n}_G$ and $|\mathbf{I}_C + \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{H}/N|^{-1}$ leads to (2.24). This completes the proof.

□

Remark 2.4.1 *If $S = 1$, (2.23) reduces to the* LRT *statistic for* MANOVA *(Mardia* et al., *1979, pp. 138-139). Hence, Theorem 2.3 generalizes* MANOVA *to the case of mixed binary and continuous data.*

Remark 2.4.2 *Conditioning the joint distribution of* $\mathbf{n}_1, \cdots, \mathbf{n}_G$ *on* $\mathbf{t}$ *follows the usual way of eliminating the dependence of the joint distribution on* $\boldsymbol{\pi}$ *(see, e. g., Read and Cressie, 1988; Koehler and Wilson, 1986). The same approach was adopted by Afifi and Elashoff (1969) in the two-sample case.*

Remark 2.4.3 *The condition* $n_{gs} > 0 \; \forall g, s$ *(i.e., each state in each population has at least 1 observation) is necessary so that all unknown parameters will be estimable. This results in a truncated conditional joint distribution for* $\mathbf{n}_1, \cdots, \mathbf{n}_G$ *given* $\mathbf{t}$.

Remark 2.4.4 *Note that (2.26) can be viewed as a generalization of the multivariate hypergeometric distribution (Bishop* et al., *1975, pp. 450-452), as it reduces to the latter in the case* $S = 2$.

Remark 2.4.5 *Properties and special cases of the Wilks'* λ*-distribution are discussed in Mardia* et al. *(1979, pp. 81-84). Approximations are found in Mardia and Zemroch (1978) and Pearson and Hartley (1972).*

Remark 2.4.6 *Under* K*, the conditional distribution of* $|\mathbf{I}_C + \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{H}/N|^{-1}$ *becomes a noncentral Wilks'* λ*-distribution (Seber, 1984, p. 42). Also, the matrix* $\mathbf{H}$ *is usually referred to as the between-groups SSP in* MANOVA.

The following corollary is obtained from Theorem 2.3 for the two-sample mixed-data case (i.e., $G = 2$).

Corollary 2.3.1 *Consider the two-sample case (i.e.,* $G = 2$*). The* LRT *statis-*

tic in (2.23) becomes

$$\lambda^{-2/N} = b^{-1}(n_{11},\cdots,n_{2S}) \times \left[1+\sum_{s=1}^{S}\frac{n_{1s}n_{2s}}{n_{\cdot s}N}(\overline{\mathbf{y}}_{1s}-\overline{\mathbf{y}}_{2s})^{\top}\widehat{\boldsymbol{\Sigma}}^{-1}(\overline{\mathbf{y}}_{1s}-\overline{\mathbf{y}}_{2s})\right], \tag{2.27}$$

where $b(n_{11},\cdots,n_{2S}) = \left[(n_{1\cdot}^{n_{1\cdot}}n_{2\cdot}^{n_{2\cdot}}/N^{N})\prod_{s=1}^{S}(n_{1s}+n_{2s})^{n_{1s}+n_{2s}}/(n_{1s}^{n_{1s}}n_{2s}^{n_{2s}})\right]^{2/N}$, *and* H *is rejected if and only if* $\lambda^{-2/N} > c_{\alpha}$, *where* c_{α} *is obtained from*

$$\alpha = \sum_{n_{11},\cdots,n_{2S}} p(n_{11},\cdots,n_{2S}\mid \mathbf{t},H) \times \Pr\left[U^{(M)} > c(n_{11},\cdots,n_{2S})\right], \tag{2.28}$$

where

$$p(n_{11},\cdots,n_{2S}\mid \mathbf{t},H) = \frac{n_{1\cdot}!n_{2\cdot}!\prod_{s=1}^{S}[(n_{1s}+n_{2s})!/n_{gs}!]}{\sum_{n_{11},\cdots,n_{2S}} n_{1\cdot}!n_{2\cdot}!\prod_{s=1}^{S}[(n_{1s}+n_{2s})!/n_{gs}!]}, \tag{2.29}$$

with the summations taken over all $\{n_{11},\cdots,n_{2S}\}$ *such that* $n_{gs} > 0\ \forall g,s$ *and* $n_{1s}+n_{2s} = n_{\cdot s}\ \forall s$, $c(n_{11},\cdots,n_{2S}) = b(n_{11},\cdots,n_{2S})c_{\alpha} - 1$, *and* $U^{(M)}$ *has the same distribution as that, under* H *and conditional on* $\{n_{11},\cdots,n_{2S}\}$, *of the sum of non-zero roots* $\xi_1,\cdots,\xi_M$ *of the following determinantal equation:*

$$\left|\sum_{s=1}^{S}\frac{n_{1s}n_{2s}}{n_{\cdot s}N}(\overline{\mathbf{y}}_{1s}-\overline{\mathbf{y}}_{2s})(\overline{\mathbf{y}}_{1s}-\overline{\mathbf{y}}_{2s})^{\top} - \xi\widehat{\boldsymbol{\Sigma}}\right| = 0.$$

Proof. First, note that

$$\left|\mathbf{I}_C + \frac{\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{H}}{N}\right| = 1+\sum_{s=1}^{S}\frac{n_{1s}n_{2s}}{n_{\cdot s}N}(\overline{\mathbf{y}}_{1s}-\overline{\mathbf{y}}_{2s})^{\top}\widehat{\boldsymbol{\Sigma}}^{-1}(\overline{\mathbf{y}}_{1s}-\overline{\mathbf{y}}_{2s}), \tag{2.30}$$

by A2.3m of Mardia *et al.* (1979, p. 458) and since $\mathbf{H}$ can be written as

$$\mathbf{H} = \sum_{s=1}^{S}\frac{n_{1s}n_{2s}}{n_{\cdot s}}(\overline{\mathbf{y}}_{1s}-\overline{\mathbf{y}}_{2s})(\overline{\mathbf{y}}_{1s}-\overline{\mathbf{y}}_{2s})^{\top}.$$

Next, assume that H holds and $\{n_{11}, \cdots, n_{2S}\}$ is fixed. It follows that $\sqrt{n_{1s}n_{2s}}(\overline{\mathbf{y}}_{1s} - \overline{\mathbf{y}}_{2s})/\sqrt{n_{.s}} \sim \mathcal{N}_C(\mathbf{0}, \boldsymbol{\Sigma})$ $\forall s$, so that $\mathbf{M}_1 = \sum_{s=1}^{S} n_{1s}n_{2s}(\overline{\mathbf{y}}_{1s} - \overline{\mathbf{y}}_{2s})(\overline{\mathbf{y}}_{1s} - \overline{\mathbf{y}}_{2s})^\top / n_{.s} \sim \mathcal{W}_C(\boldsymbol{\Sigma}, S)$, independently of $\mathbf{M}_2 = N\widehat{\boldsymbol{\Sigma}} \sim \mathcal{W}_C(\boldsymbol{\Sigma}, N - SG)$. The rest of the proof parallels that of Theorem 2.2, with $\xi_1, \cdots, \xi_M$ the non-zero roots of

$$\begin{aligned} 0 &= \left|\mathbf{M}_1\mathbf{M}_2^{-1} - \xi\mathbf{I}_C\right| \\ &= \left|\sum_{s=1}^{S} \frac{n_{1s}n_{2s}}{n_{.s}N}(\overline{\mathbf{y}}_{1s} - \overline{\mathbf{y}}_{2s})(\overline{\mathbf{y}}_{1s} - \overline{\mathbf{y}}_{2s})^\top - \xi\widehat{\boldsymbol{\Sigma}}\right|. \end{aligned}$$

□

The following corollary for the case $G > 2, C = 1$ and $S = 2$ is obtained by applying Theorem 2.3 to the model considered in Example 2.1.

Corollary 2.3.2 *Consider the model in Example 2.1. The* LRT *statistic in the case where $G > 2$ is given by*

$$\begin{aligned} \lambda^{-2/N} &= b^{-1}(n_1, \cdots, n_G) \\ &\quad \left\{1 + \sum_{g=1}^{G}\left[\frac{n_g}{N\widehat{\sigma}^2}(\overline{Y}_{g1} - \overline{Y}_{\cdot 1})^2 + \frac{N_g - n_g}{N\widehat{\sigma}^2}(\overline{Y}_{g2} - \overline{Y}_{\cdot 2})^2\right]\right\}, \end{aligned} \tag{2.31}$$

where $b(n_1, \cdots, n_G) = \left[\widehat{p}^n\widehat{q}^{N-n} / \left(\prod_{g=1}^{G} \widehat{p}_g^{n_g}\widehat{q}_g^{N_g - n_g}\right)\right]^{2/N}$,

$$N\widehat{\sigma}^2 = \sum_{g=1}^{G}\left[n_g S_{g1}^2 + (N_g - n_g)S_{g2}^2\right], \tag{2.32}$$

$\widehat{p} = n/N$, $\widehat{p}_g = n_g/N_g$, $\widehat{q} = 1 - \widehat{p}$, $\widehat{q}_g = 1 - \widehat{p}_g$, $\overline{Y}_{gs}$ and S_{gs}^2 are the respective sample state mean and variance (uncorrected for bias) of observations belonging to state s from population g, $n = \sum_{g=1}^{G} n_g$, $N = \sum_{g=1}^{G} N_g$, with n_g and $N_g - n_g$ the numbers of observations belonging to states 1 and 2, respectively,

from population $g = 1, \cdots, G$. Note that it is assumed in (2.32) that, without loss of generality, the first n_g observations from population g belong to state 1. The α-critical value c_α is obtained from

$$\alpha = \sum_{n_1, \cdots, n_G} p(n_1, \cdots, n_G \mid n, H) \Pr\left[F_{d_1, d_2} > c(n_1, \cdots, n_G)\right], \tag{2.33}$$

where

$$p(n_1, \cdots, n_G \mid n, H) = \frac{\prod_{g=1}^{G} \binom{N_g}{n_g}}{\sum_{n_1, \cdots, n_G} \prod_{g=1}^{G} \binom{N_g}{n_g}}, \tag{2.34}$$

with the summations taken over all $\{n_1, \cdots, n_G\}$ such that $1 \leq n_g \leq N_g - 1$ $\forall g$ and $\sum_{g=1}^{G} n_g = n$, $d_1 = 2(G-1)$, $d_2 = N - 2G$, and $c(n_1, \cdots, n_G) = d_1 \left[b(n_1, \cdots, n_G) c_\alpha - 1\right] / d_2$.

The power of the test at $(p_1, \mu_{11}, \mu_{12})^\top, \cdots, (p_G, \mu_{G1}, \mu_{G2})^\top$ is given by

$$\sum_{n_1, \cdots, n_G} p(n_1, \cdots, n_G \mid n) \Pr\left[F_{d_1, d_2}^{\Delta(n_1, \cdots, n_S)} > c(n_1, \cdots, n_G)\right], \tag{2.35}$$

where

$$p(n_1, \cdots, n_G \mid n) = \frac{\prod_{g=1}^{G} \rho_g^{n_g} \binom{N_g}{n_g}}{\sum_{n_1, \cdots, n_G} \prod_{g=1}^{G} \rho_g^{n_g} \binom{N_g}{n_g}}, \tag{2.36}$$

$\Delta(n_1, \cdots, n_G) = \sum_{g=1}^{G} [n_g(\mu_{g1} - \overline{\mu}_{\cdot 1})^2 + (N_g - n_g)(\mu_{g2} - \overline{\mu}_{\cdot 2})^2] / \sigma^2$, $\overline{\mu}_{\cdot s} = \sum_{g=1}^{G} n_{gs} \mu_{gs} / n_{\cdot s}$, and $p_g = \delta \rho_g / (1 + \delta \rho_g)$, with the scale factor δ adjusted so that $\sum_{g=1}^{G} N_g p_g = n$.

Proof. The test statistic in (2.31) follows immediately from Theorem 2.3 by taking $C = 1$ and $S = 2$.

The null distribution in (2.33) is derived by noting that $\sum_{g=1}^{G} n_g (\overline{Y}_{g1} - \overline{Y}_{\cdot 1})^2 / \sigma^2$, $\sum_{g=1}^{G} (N_g - n_g)(\overline{Y}_{g2} - \overline{Y}_{\cdot 2})^2 / \sigma^2$, and $N\widehat{\sigma}^2 / \sigma^2$ are independent χ^2_{G-1},

χ^2_{G-1}, and χ^2_{N-2G}, respectively. Hence,

$$\begin{aligned}\sum_{g=1}^{G}\left[\frac{n_g}{N\widehat{\sigma}^2}(\overline{Y}_{g1}-\overline{Y}_{\cdot 1})^2+\frac{N_g-n_g}{N\widehat{\sigma}^2}(\overline{Y}_{g2}-\overline{Y}_{\cdot 2})^2\right] &= \frac{2(G-1)}{N-2G} \\ &\quad\times\frac{\chi^2_{2(G-1)}/(2G-2)}{\chi^2_{N-2G}/(N-2G)} \\ &\sim \frac{d_1}{d_2}F_{d_1,d_2},\end{aligned}$$

under H and conditional on $\{n_1,\cdots,n_G\}$. The rest of (2.33) uses the fact that $n_1,\cdots,n_G$, conditional on $\sum_{g=1}^{G}n_g=n$, is multivariate hypergeometric with parameters n, N and $(N_1,\cdots,N_G)$, under H (Bishop *et al.*, 1975, p. 452).

The power of the test immediately follows from the fact that $\sum_{g=1}^{G}n_g(\overline{Y}_{g1}-\overline{Y}_{\cdot 1})^2/\sigma^2$ and $\sum_{g=1}^{G}(N_g-n_g)(\overline{Y}_{g2}-\overline{Y}_{\cdot 2})^2/\sigma^2$ are independent noncentral χ^2_{G-1} with noncentrality parameters $\sum_{g=1}^{G}n_g(\mu_{g1}-\overline{\mu}_{\cdot 1})^2/\sigma^2$ and $\sum_{g=1}^{G}(N_g-n_g)(\mu_{g2}-\overline{\mu}_{\cdot 2})^2/\sigma^2$, respectively, and $(n_1,\cdots,n_G)^\top$, given $\sum_{g=1}^{G}n_g=n$, has a multivariate extended hypergeometric distribution with parameters n, G, $(N_1,\cdots,N_G)$, and $(\rho_1,\cdots,\rho_G)$ (Harkness, 1965). This completes the proof.

□

The following corollary is obtained from Corollary 2.3.2 with $G=2$.

Corollary 2.3.3 *Consider the model in Example 2.1. The* LRT *statistic in the case where $G=2$ is given by*

$$\begin{aligned}\lambda^{-2/N} &= b^{-1}(n_1,n_2) \\ &\quad\times\left[1+\frac{n_1n_2}{nN\widehat{\sigma}^2}(\overline{Y}_{11}-\overline{Y}_{21})^2+\frac{(N_1-n_1)(N_2-n_2)}{(N-n)N\widehat{\sigma}^2}(\overline{Y}_{12}-\overline{Y}_{22})^2\right],\end{aligned} \tag{2.37}$$

where $b(n_1, n_2)$ *and* $\widehat{\sigma}^2$ *are as defined in Corollary 2.3.2. The* α*-critical value* c_α *is obtained from*

$$\alpha = \sum_{n_1=1}^{N_1-1} \frac{\binom{N_1}{n_1}\binom{N-N_1}{n-n_1}}{\binom{N}{n} - \binom{N-N_1}{n-N_1} - \binom{N-N_1}{n}} \Pr\left[F_{2,N-4} > c(n_1, n_2)\right], \tag{2.38}$$

where $c(n_1, n_2) = (N-4)[b(n_1, n_2)c_\alpha - 1]/2$, *provided* $n + N_1 \geq N$ *and* $n \geq N_1$.

The power of the test at $(p_1, \mu_{11}, \mu_{12})^\top \neq (p_2, \mu_{21}, \mu_{22})^\top$ *is given by*

$$\sum_{n_1=1}^{N_1-1} p(n_1, n_2 \mid n, p_1, p_2) \Pr\left[F_{2,N-4}^{\Delta(n_1,n_2)} > c(n_1, n_2)\right], \tag{2.39}$$

where

$$p(n_1, n_2 \mid n, p_1, p_2) = \frac{\rho^{n_1}\binom{N_1}{n_1}\binom{N-N_1}{n-n_1}}{\sum_{n_1=0}^{N_1} \rho^{n_1}\binom{N_1}{n_1}\binom{N-N_1}{n-n_1} - \rho^{N_1}\binom{N-N_1}{n-N_1} - \binom{N-N_1}{n}}, \tag{2.40}$$

with $\rho = p_1 q_2/(p_2 q_1)$ *and* $\Delta(n_1, n_2) = [n_1 n_2(\mu_{11} - \mu_{21})^2/n + (N_1 - n_1)(N_2 - n_2)(\mu_{12} - \mu_{22})^2/n]/\sigma^2$.

Proof. The proofs of (2.37)-(2.38) follow immediately from Corollary 2.3.2 by taking $G = 2$, and by noting that n_1, under H and conditional on $n_1 + n_2 = n$, is hypergeometric with parameters N_1 and $N_2 = N - N_1$ (Bishop *et al.*, 1975, p. 450), so that the normalizing constant is $1 - \left[\binom{N-N_1}{n-N_1} + \binom{N-N_1}{n}\right] / \binom{N}{n}$.

Expression (2.39) is proved using the fact that n_1, under K and conditional on $n_1 + n_2 = n$, has a noncentral (or extended) hypergeometric distribution with noncentrality parameter $\rho = p_1 q_2/(p_2 q_1)$ (Johnson *et al.*, 1992, pp. 279-282), so that the normalizing constant becomes

$$1 - \frac{\rho^{N_1}\binom{N-N_1}{n-N_1} + \binom{N-N_1}{n}}{\sum_{n_1=0}^{N_1} \rho^{n_1}\binom{N_1}{n_1}\binom{N-N_1}{n-n_1}}.$$

□

Remark 2.4.7 *If $S = 1$ in Corollary 2.3.1, the* LRT *reduces to that for testing equality of the means of two independent multivariate normal samples (Mardia* et al., *1979, pp. 139-140), or equivalently, the two-sample Hotelling T^2 test statistic given by*

$$\begin{aligned}\frac{n_1 n_2 d_2}{n^2 d_1}(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2) &= \frac{n_1 n_2 d_2}{n(n-2)d_1}(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)^\top \mathbf{S}_u^{-1}(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2) \\ &\sim F_{d_1, d_2},\end{aligned}$$

under H, where $d_1 = C$, $d_2 = n - C - 1$, and $\mathbf{S}_u = (n-2)\widehat{\boldsymbol{\Sigma}}/n$ (Mardia et al., *1979, p. 76-77). Thus, Corollary 2.3.1 generalizes the multivariate two-sample location problem to the case of mixed binary and continuous data.*

Remark 2.4.8 *The* LRT *in Corollary 2.3.1 appears in Afifi and Elashoff (1969) and is analogous to a test derived by Bar-Hen and Daudin (1995) based on a generalization of the Mahalanobis distance via the Kullback-Leibler divergence (Kullback, 1968, pp. 6-7).*

Remark 2.4.9 *If $S = 1$ in Corollary 2.3.2, the* LRT *reduces to that for testing equality of the means of several independent normal samples, or equivalently, the analysis of variance* (ANOVA) *F-ratio test (Mood* et al., *1974, pp. 435-438). Thus, Corollary 2.3.2 generalizes the univariate* ANOVA *problem to the case of mixed binary and continuous data.*

Remark 2.4.10 *If $S = 1$ in Corollary 2.3.3, the* LRT *reduces to that for testing equality of the means of two independent normal samples, or equivalently, the two-sample t-test (Mood* et al., *1974, pp. 432-435). Thus, Corollary 2.3.3*

generalizes the univariate two-sample location problem to the case of mixed binary and continuous data.

2.4.2 Heterogeneous Cases

The Behrens-Fisher problem originally arose in the test of equality of means of two independent univariate normal samples ($G = 2$) when the variances are not equal. It has since been used to refer to the analogous problem in the case of $G > 2$ multivariate normal samples. Two situations where such problem can arise in the mixed-data setting are now briefly studied. The first only assumes *within-population homogeneity*, i.e., $\boldsymbol{\Sigma}_{g1} = \cdots = \boldsymbol{\Sigma}_{gS} = \boldsymbol{\Sigma}_g$ for $g = 1, \cdots, G$. The second, on the other hand, only assumes *within-state homogeneity*, i.e., $\boldsymbol{\Sigma}_{1s} = \cdots = \boldsymbol{\Sigma}_{Gs} = \boldsymbol{\Sigma}_{\cdot s}$ for $s = 1, \cdots, S$. Note that complete homogeneity in § 2.4.1 is simply a special case of these.

Consider the situation where within-state homogeneity holds. The log-likelihood $\ell^{(C)}$ of the continuous data can then be written, in this case, as

$$\begin{aligned} \ell^{(C)} &= -\sum_{g=1}^{G}\sum_{s=1}^{S}\left\{\frac{n_{gs}}{2}\log\mid 2\pi\boldsymbol{\Sigma}_{\cdot s}\mid +\frac{n_{gs}}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_{\cdot s}^{-1}\left(\mathbf{S}_{gs}+\mathbf{d}_{gs}\mathbf{d}_{gs}^{\top}\right)\right]\right\} \quad (2.41)\\ &= -\sum_{s=1}^{S}\left\{\frac{n_{\cdot s}}{2}\log\mid 2\pi\boldsymbol{\Sigma}_{\cdot s}\mid +\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}_{\cdot s}^{-1}\sum_{g=1}^{G}n_{gs}\left(\mathbf{S}_{gs}+\mathbf{d}_{gs}\mathbf{d}_{gs}^{\top}\right)\right]\right\}, \end{aligned}$$

where $n_{\cdot s} = \sum_{g=1}^{G} n_{gs}$. That for the binary part, $\ell^{(D)}$, follows from (2.20).

Under the alternative hypothesis K, the MLE $\widehat{\boldsymbol{\theta}}_g$ of $\boldsymbol{\theta}_g$, $g = 1, \cdots, G$, is as given in (2.21) and that for $\boldsymbol{\Sigma}_{\cdot s}$ can be shown as

$$n_{\cdot s}\widehat{\boldsymbol{\Sigma}}_{\cdot s} = \sum_{g=1}^{G} n_{gs}\mathbf{S}_{gs},$$

for $s = 1, \cdots, S$.

Similarly, under the null hypothesis H where $\boldsymbol{\pi}_1 = \cdots = \boldsymbol{\pi}_G = \boldsymbol{\pi}$ and $\boldsymbol{\mu}_{1s} = \cdots = \boldsymbol{\mu}_{Gs} = \boldsymbol{\mu}_{\cdot s}$, the (restricted) MLEs of $\boldsymbol{\pi}$ and $\boldsymbol{\mu}_{\cdot s}$ are as given in (2.22), and that for $\boldsymbol{\Sigma}_{\cdot s}$ is given by

$$n_{\cdot s}\hat{\hat{\boldsymbol{\Sigma}}}_{\cdot s} = \sum_{g=1}^{G}\sum_{i=1}^{n_{gs}}(\mathbf{y}_{gsi} - \overline{\mathbf{y}}_{\cdot s})(\mathbf{y}_{gsi} - \overline{\mathbf{y}}_{\cdot s})^{\top}.$$

These MLEs can be explained by viewing the problem as consisting of S G-sample homogeneous tests as in § 2.4.1.

The LRT statistic is then

$$\begin{aligned}\lambda &= \prod_{g=1}^{G}\prod_{s=1}^{S}\frac{n_{g\cdot}^{n_{g\cdot}}n_{\cdot s}^{n_{gs}}}{N^{n_{g\cdot}}n_{gs}^{n_{gs}}} \times \prod_{s=1}^{S}\left(\frac{|\hat{\hat{\boldsymbol{\Sigma}}}_{\cdot s}|}{|\hat{\boldsymbol{\Sigma}}_{\cdot s}|}\right)^{-n_{\cdot s}/2} \qquad (2.42)\\ &= b^{N/2}(n_{11},\cdots,n_{GS}) \\ &\quad \times \prod_{s=1}^{S}\left[1 + \sum_{g=1}^{G}\frac{n_{gs}}{n_{\cdot s}}(\overline{\mathbf{y}}_{gs} - \overline{\mathbf{y}}_{\cdot s})^{\top}\hat{\boldsymbol{\Sigma}}_{\cdot s}^{-1}(\overline{\mathbf{y}}_{gs} - \overline{\mathbf{y}}_{\cdot s})\right]^{-n_{\cdot s}/2},\end{aligned}$$

where $b(n_{11},\cdots,n_{GS})$ is as defined in Theorem 2.3. Equivalently,

$$-2\log\lambda = -N\log b(n_{11},\cdots,n_{GS}) + \sum_{s=1}^{S} n_{\cdot s}\log V(n_{\cdot s}), \qquad (2.43)$$

where, conditional on $n_{11},\cdots,n_{GS}$, $V(n_{\cdot 1}),\cdots,V(n_{\cdot S})$ are independent such that

$$\frac{n_{\cdot s} - G - C + 1}{CG}[V(n_{\cdot s}) - 1] \sim F_{CG,n_{\cdot s}-G-C+1}.$$

The same approach as before can then be employed to get the exact distribution of (2.43), except that it now involves neither the χ^2 nor F distributions but rather the distribution of a sum of transformations of independent F random variables.

It should be mentioned that an analogous situation in MANOVA arose in an earlier paper by Geisser (1963), where a *uniform* or *compound symmetric* structure for the covariance matrix (i.e., equal variances and equal covariances) is assumed for the data. In the one-sample multivariate normal case, the LRT leads to a test statistic very similar to (2.42) and (2.43) (see Eq. (1.5) of Geisser, 1963). Approximations to distributions of linear combinations of independent F (and beta) random variables are given by Jóhannesson and Giri (1995), Dyer (1982) and Morrison (1971).

The case of within-population homogeneity is where the Behrens-Fisher problem arises. In what follows, it is assumed that $G = 2$ for simplicity. In this case, the same MLEs under K as those derived previously for the case of within-state homogeneity still obtains except for

$$\begin{aligned} n_{g\cdot}\widehat{\boldsymbol{\Sigma}}_g &= \sum_{s=1}^{S} n_{gs}\mathbf{S}_{gs} \\ &\equiv n_{g\cdot}\mathbf{S}_{g\cdot}, \end{aligned}$$

for $g = 1, 2$. Derivation of the (restricted) MLEs of $\boldsymbol{\mu}_{\cdot s}$ and $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ follows from standard results (Mardia *et al.*, 1979, pp. 103-105) and are given by

$$\begin{aligned} \widehat{\widehat{\boldsymbol{\mu}}}_{\cdot s} &= \left[n_{1s}\widehat{\widehat{\boldsymbol{\Sigma}}}_1 + n_{2s}\widehat{\widehat{\boldsymbol{\Sigma}}}_2\right]^{-1}\left[n_{1s}\widehat{\widehat{\boldsymbol{\Sigma}}}_1\overline{\mathbf{y}}_{1s} + n_{2s}\widehat{\widehat{\boldsymbol{\Sigma}}}_2\overline{\mathbf{y}}_{2s}\right] \\ n_{g\cdot}\widehat{\widehat{\boldsymbol{\Sigma}}}_g &= n_{g\cdot}\mathbf{S}_{g\cdot} + \sum_{s=1}^{S} n_{gs}(\overline{\mathbf{y}}_{gs} - \widehat{\widehat{\boldsymbol{\mu}}}_{\cdot s})(\overline{\mathbf{y}}_{gs} - \widehat{\widehat{\boldsymbol{\mu}}}_{\cdot s})^{\top}, \end{aligned}$$

for $g = 1, 2; s = 1, \cdots, C$. These estimators need to be calculated iteratively as suggested in Mardia *et al.* (1979, pp. 142-143). The asymptotic distribution of the LRT statistic (Ogawa *et al.*, 1957) may then be used to carry out a test of (2.19). It is an open question whether solutions to the Behrens-Fisher

problem (Hussein and Carrière, 2001; Jordan and Krishnamoorthy, 1995) can be adapted to the mixed-data setting.

2.5 Discussion

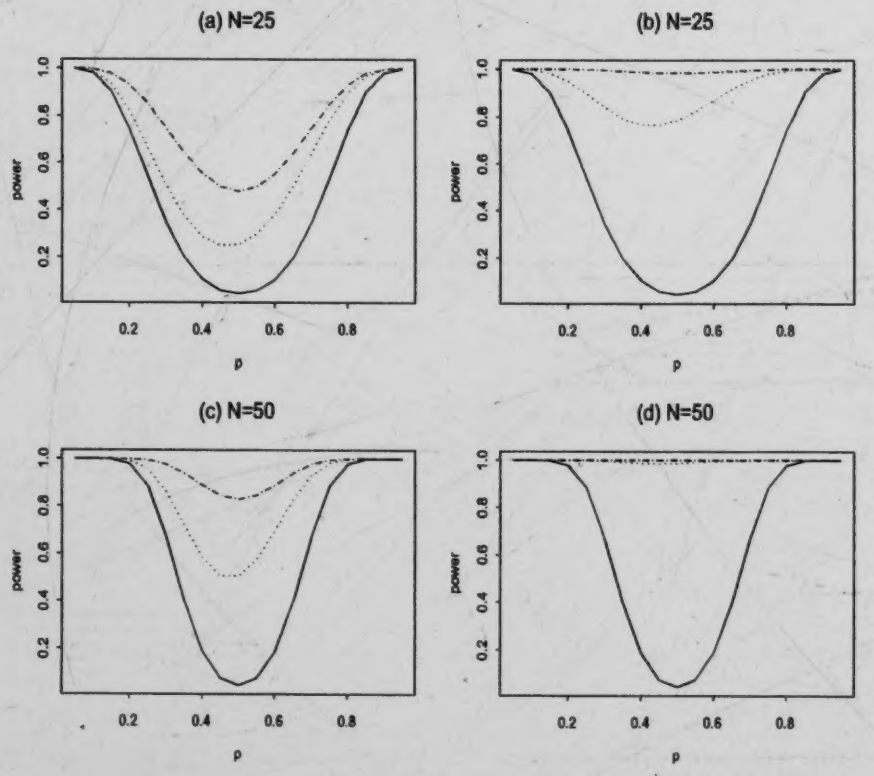
This chapter was concerned with tests of location hypotheses for mixed multivariate data distributed according to the general location model. Likelihood ratio tests for the one-sample and multi-sample problems were obtained and their exact distributions were derived. These LRTs provide global tests of location hypotheses and thus avoid the problem of multiple testing. The test statistics are similar to their continuous case counterparts and are simple and easy to calculate. In addition, critical values of the tests can be easily calculated by conventional methods (see Table 2.1).

The likelihood ratio approach was employed to construct global tests of mixed data location hypotheses because it allows for a general non-*ad hoc* approach of simultaneously accounting for both the discrete (i.e., multinomial) and continuous variables in the data. The approach parallels that of Afifi and Elashoff (1969) and is an alternative to the dissimilarity-based tests proposed by Morales *et al.* (1998). These tests, it should be noted, are all asymptotic, unlike the exact LRTs derived in this chapter. By modelling the joint distribution of the mixed variables as a general location model, the resulting LRTs can be viewed as extensions of classical LRTs in the one-sample and multi-sample problems based on normal distributions. The power functions of the LRTs were also obtained and investigated, in particular, for the one-sample

case with one binary and one continuous variable. The simulations indicate that the LRT outperforms the *separate test* approach to a considerable extent.

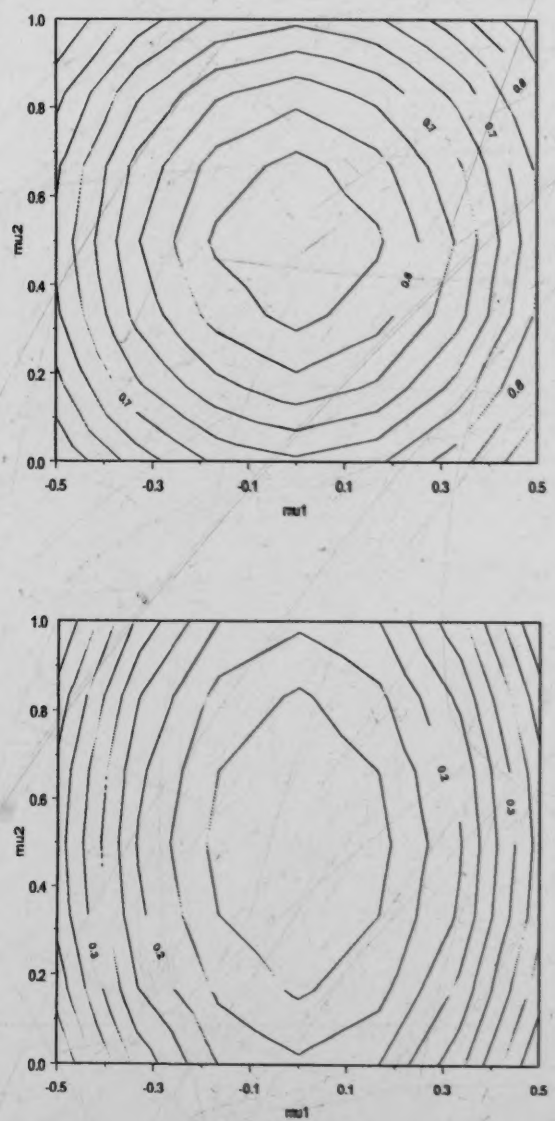
Finally, the Behrens-Fisher problem as it arises in the mixed-data setting was briefly discussed. The same complications that it engenders in the (strictly) continuous case were also identified in the mixed-data case. The dissimilarity-based tests proposed by Morales *et al.* (1998) could be applied to this case, as the tests were derived without the assumption of complete homogeneity. However, they rely on the asymptotic distributions of the test statistics and their performance, especially in small samples, has yet to be studied. It is still an open question whether solutions (see, e.g., Hussein and Carrière, 2001, for a recent survey) to the Behrens-Fisher problem can be adapted to the case of mixed binary and continuous data.

Figure 2.1: *Plots of the Power Function of* LRT *with* $C = 1, S = 2$ *and unknown* σ^2*, for* $p_0 = 0.5$ *and fixed State Means.*



NOTE: *Displayed are plots of the power function in Corollary 2.2.1 at* $\alpha = 0.05, \sigma^2 = 1$. *For (a) and (c), the solid lines correspond to* $\mu_1 = \mu_{01} = 0, \mu_2 = \mu_{02} = 0.5$*; dotted lines to* $\mu_1 = 0, \mu_{01} = 0.5, \mu_2 = \mu_{02} = 0.5$*; dashed lines to* $\mu_1 = \mu_{02} = 0, \mu_2 = \mu_{02} = 0.5$. *For (b) and (d), the solid lines correspond to* $\mu_1 = \mu_{01} = 0, \mu_2 = \mu_{02} = 1$*; dotted lines to* $\mu_1 = 0, \mu_{01} = 1, \mu_2 = \mu_{02} = 1$*; dashed lines to* $\mu_1 = \mu_{02} = 0, \mu_2 = \mu_{02} = 1$.

Figure 2.2: *Contour Plots of the Power Function of* LRT *with* $C = 1, S = 2$ *and unknown* σ^2*, for* $N = 25$ *and* $p = 0.5$.



NOTE: *Displayed are contour plots of the power function in Corollary 2.2.1 at* $\alpha = 0.05$*, with* $(\mu_{01}, \mu_{02}) = (0, 0.5), \sigma^2 = 1, N = 25$*, and* $p_0 = 0.25$ *(top),* 0.5 *(bottom). Note that* H *is false for* $p_0 = 0.25$ *(top) while it is true for* $p_0 = 0.5$ *(bottom) at* $(\mu_1, \mu_2) = (0, 0.5)$.

Table 2.1: *Critical Values* c_α *for the* LRT *with* $C = 1, S = 2$ *and unknown* σ^2.

		N					
p_0	α	10	15	20	25	30	50
0.05	0.01	3.798619	2.2596	1.791351	1.576424	1.453066	1.246445
	0.05	2.496786	1.741778	1.483547	1.360345	1.286645	1.161088
0.1	0.01	3.611616	2.211619	1.781555	1.577642	1.459268	1.255812
	0.05	2.381277	1.708373	1.477761	1.362857	1.294167	1.171379
0.25	0.01	3.626509	2.260955	1.825464	1.614376	1.489984	1.264061
	0.05	2.40079	1.752161	1.518428	1.397828	1.316231	1.175092
0.5	0.01	3.744688	2.327917	1.848998	1.622168	1.490199	1.263739
	0.05	2.504224	1.791348	1.527104	1.394488	1.315474	1.175038

NOTE: *Shown are critical values for the likelihood ratio test in Corollary 2.2.1 (unknown* σ^2*) at level* $\alpha = .01, .05$*, with* $p_0 = .05, .1, .25, .5$.

Table 2.2: *Power Comparison of the* LRT *with* $C = 1, S = 2$ *and unknown* σ^2 *against the Separate Test Approach, for* $H : \boldsymbol{\theta} = (.3, 50, 25)^\top$ *based on 10,000 Monte Carlo samples of sizes* $N = 15$ *and 25.*

		N=15		N=25	
Case	α	LRT	ST	LRT	ST
(0)	0.05	0.05	0.0362	0.05	0.0429
	0.01	0.01	0.0093	0.01	0.0071
(a)	0.05	0.0626	0.0443	0.0656	0.0555
	0.01	0.0138	0.0089	0.0149	0.0096
(b)	0.05	0.3005	0.2358	0.5001	0.4258
	0.01	0.1142	0.0805	0.2484	0.1822
(c)	0.05	0.5974	0.4947	0.8499	0.7824
	0.01	0.3377	0.2384	0.6469	0.5243
(d)	0.05	0.8578	0.7533	0.9872	0.9664
	0.01	0.6199	0.4543	0.9306	0.8322

NOTE: *Shown are the powers of the likelihood ratio test* (LRT) *in Corollary 2.2.1 against those (empirical) of the separate test* (ST) *for* (0): $\boldsymbol{\theta} = (.3, 50, 25)^\top$; (a): $\boldsymbol{\theta} = (.35, 50, 25)^\top$; (b): $\boldsymbol{\theta} = (.35, 52.5, 22.5)^\top$; (c): $\boldsymbol{\theta} = (.4, 55, 22.5)^\top$; *and* (d): $\boldsymbol{\theta} = (.4, 55, 20)^\top$, *with* $\sigma^2 = 25$. *Note that* $\rho_{\mathbf{x},Y}$ *is* .916 *in* (0), .922 *in* (a), .944 *in* (b), .954 *in* (c), *and* .96 *in* (d).

Table 2.3: *Power Comparison of the* LRT *with* $C = 1, S = 2$ *and unknown* σ^2 *against the Separate Test Approach, for* $H : \boldsymbol{\theta} = (.5, 50, 45)^\top$ *based on 10,000 Monte Carlo samples of sizes* $N = 15$ *and 25.*

		N=15		N=25	
Case	α	LRT	ST	LRT	ST
(0)'	0.05	0.05	0.042	0.05	0.0395
	0.01	0.01	0.008	0.01	0.0077
(a)'	0.05	0.179	0.1507	0.2989	0.2747
	0.01	0.0579	0.0482	0.1215	0.1161
(b)'	0.05	0.2862	0.228	0.4982	0.412
	0.01	0.1044	0.0748	0.2467	0.1844
(c)'	0.05	0.2862	0.2338	0.4982	0.4163
	0.01	0.1044	0.0762	0.2467	0.1804
(d)'	0.05	0.6365	0.5982	0.898	0.873
	0.01	0.3547	0.3207	0.7144	0.6766

NOTE: *Shown are the powers of the likelihood ratio test* (LRT) *in Corollary 2.2.1 against those (empirical) of the separate test* (ST) *for* (0)': $\boldsymbol{\theta} = (.5, 50, 45)^\top$; (a)': $\boldsymbol{\theta} = (.45, 50, 42.5)^\top$; (b)': $\boldsymbol{\theta} = (.45, 52.5, 42.5)^\top$; (c)': $\boldsymbol{\theta} = (.55, 52.5, 42.5)^\top$; *and* (d)': $\boldsymbol{\theta} = (.55, 55, 47.5)^\top$, *with* $\sigma^2 = 25$. *Note that* $\rho_{\mathbf{x},Y}$ *is* .447 *in* (0)', .598 *in* (a)' *and* (d)', *and* .705 *in* (b)' *and* (c)'.

Chapter 3

Pairwise Likelihood Approach to Grouped Continuous Model and Its Extension

3.1 Introduction

Precise measurement of study variables is not always possible in practice. Because of this limitation, researchers, especially in the medical and social sciences, rely on using ordinal instead of interval (or scale) variables in their studies, as when a patient's state of health is evaluated as, say, very poor, poor, average, good, or very good, in the absence of a more precise measure of the patient's condition.

A common approach to handling ordinal data is to assume that the ordinal variables are *coarsely* measured versions of unobservable continuous variables called *latent variables*, and are obtained by partitioning or *thresholding* the space of the latent variables into non-overlapping intervals. Pearson (1904) was the first to adopt this approach, and his work has since been extended in several directions (Anderson and Philips, 1981; McCullagh, 1980).

This chapter is concerned with the *grouped continuous model*, a model for multivariate ordinal data that assumes a multivariate normal distribution for the latent variables, leading to a *probit model* for the ordinal variables. It was introduced by Anderson and Pemberton (1985) as a generalization of the corresponding univariate model developed earlier by Anderson and Philips (1981) and McCullagh (1980). It relies on the so-called *polychoric correlation* (Drasgow, 1986) to model the covariance structure of the data, in contrast to a log-linear model (Agresti 1990; 1984) that relies on *odds ratio* (also called *cross-product ratio*) or the *Pearson's correlation* as a measure of ordinal data association (Bishop *et al.*, 1975, pp. 376-393). Unlike the Pearson's correlation, polychoric correlation does not restrict the correlation parameter space. Moreover, the number of polychoric correlations does not increase with the number of levels that the ordinal data can assume, a common problem with using odds ratios.

Another advantage of the grouped continuous model is that it can easily be extended to mixed data with ordinal and continuous variables. Such extension of the model, called the *conditional grouped continuous model*, was introduced by Anderson and Pemberton (1985) in the context of regression analysis of multivariate ordinal outcomes, where the continuous variables were treated as covariates. It was later studied by Poon and Lee (1987; 1986) as a model for mixed data, with both continuous and ordinal variables considered as outcomes. They investigated maximum likelihood estimation of polychoric and *polyserial* correlations, with the latter representing the correlations be-

tween the continuous and ordinal variables (Drasgow, 1986). The extension of the model to the multi-sample case was investigated by Poon and Lee (1992). See also Ronning and Kukuk (1996).

Maximum likelihood estimation of the polychoric correlations and the *cutpoints* (or thresholds) in the grouped continuous model has been previously studied by Tallis (1962) and Martinson and Hamdan (1971). The first tackled the problem of maximum likelihood estimation in the special case of two ordinal variables, each with three levels. The second generalized Tallis's (1962) approach to bivariate ordinal data with arbitrary number of levels and developed a two-step estimation method for the parameters (see also Lee, 1985; Olsson, 1979). Anderson and Pemberton (1985) proposed a computationally feasible method in the general case, which consists in first estimating the cutpoints marginally, and then estimating the polychoric correlations based on the likelihood with the cutpoints replaced by their estimates. For a survey of estimation methods for polychoric correlations and algorithms for implementing them, see Drasgow (1986). More recent references include Poon *et al.* (1990), Lee *et al.* (1989), and Lee and Lau (1986).

The corresponding estimation problem for polyserial correlations was first studied by Tate (1955; 1954) in the case of a single dichotomous variable, where the polyserial correlation is known as the *point-biserial correlation.* His results were later extended in various ways by Lee and Poon (1986), Olsson *et al.* (1982), Cox (1974), and Hannan and Tate (1965).

The combined problem of estimating polychoric and polyserial correla-

tions arising from the conditional grouped continuous model was discussed by Poon and Lee (1987). Besides maximum likelihood estimation, they proposed an alternative, now known as the *partition maximum likelihood* (PML) method, which entails partitioning the model into sub-models and then averaging the estimates obtained from these sub-models. Poon *et al.* (1990) applied this method to the multi-sample case and derived the asymptotic distribution of the PML estimates. The *pairwise* PML method, which only considers pairwise (or bivariate) sub-models, was recently proposed by Bedrick *et al.* (2000) and Lapidus (1998). Although these PML methods are less computationally demanding than the maximum likelihood approach, estimation of the parameters is done separately for several models with common parameters, and hence, the efficiency of the estimates may be compromised. As well, because they are non-simultaneous, they yield multiple sets of estimates with no clear prescription for combining them to obtain the final estimates. Poon and Lee (1987), and others have suggested simply averaging the estimates, clearly an ad-hoc solution. There is thus a need for a more systematic estimation method for the model than the PML methods.

The grouped continuous model is formally introduced in the next section. In § 3.3, maximum likelihood estimation for the model is briefly reviewed. An alternative method based on the *pairwise likelihood approach* (Kuk and Nott, 2000; Nott and Rydén, 1999) is detailed in § 3.4. Consistency and asymptotic normality of the estimates are proved in § 3.5 and are used to construct large-sample tests of hypotheses. A simulation study of the efficiency

and bias of *maximum pairwise likelihood estimates* is reported in § 3.6. The corresponding development for the conditional grouped continuous model is given in § 3.7. Finally, the chapter concludes with a discussion in § 3.8.

3.2 Grouped Continuous Model

Suppose $\mathbf{z} = (Z_1, \cdots, Z_Q)^\top$ is a vector of ordinal variables such that Z_q has $L_q + 1$ levels $a_q^1 < \cdots < a_q^{L_q+1}$, $q = 1, \cdots, Q$. Underlying $\mathbf{z}$ is $\mathbf{y}^* = (Y_1^*, \cdots, Y_Q^*)^\top$, a vector of continuous latent variables whose relationship with $\mathbf{z}$ is defined by the following *threshold model*:

$$\begin{aligned} Z_q = a_q^1 \quad &\Longleftrightarrow \quad -\infty < Y_q^* \leq \alpha_q^1, \\ Z_q = a_q^{\ell_q} \quad &\Longleftrightarrow \quad \alpha_q^{\ell_q - 1} < Y_q^* \leq \alpha_q^{\ell_q}, \quad (\ell_q = 2, \cdots, L_q) \\ Z_q = a_q^{L_q+1} \quad &\Longleftrightarrow \quad \alpha_q^{L_q} < Y_q^* < +\infty, \end{aligned} \tag{3.1}$$

where $\{\alpha_q^0 = -\infty, \alpha_q^1, \cdots, \alpha_q^{L_q}, \alpha_q^{L_q+1} = +\infty\}$ are the unknown cutpoints or thresholds. Without loss of generality, it is assumed that $a_q^{\ell_q} = \ell_q$, $\ell_q = 1, \cdots, L_q + 1$. Note that $\mathbf{z}$ defines a $(L_1 + 1) \times \cdots \times (L_Q + 1)$ contingency table.

Assuming that $\mathbf{y}^* \sim \mathcal{N}_Q(\mathbf{0}, \mathbf{R})$, with $\mathbf{R}$ a correlation matrix, the *grouped continuous model* for $\mathbf{z}$ may be defined as follows.

Definition 3.1 *The vector* $\mathbf{z}$ *is said to be distributed according to the grouped continuous model if and only if*

$$\begin{aligned} \Pr(\mathbf{z} = \boldsymbol{\ell}) &= \Pr(Z_1 = \ell_1, \cdots, Z_Q = \ell_Q) \\ &= \int_S \phi_Q(\mathbf{v} \mid \mathbf{R}) \mathrm{d}\mathbf{v}, \end{aligned} \tag{3.2}$$

where $\mathcal{S} = \{(v_1, \cdots, v_Q) : \alpha_q^{\ell_q - 1} < v_q \leq \alpha_q^{\ell_q}, q = 1, \cdots, Q\}$ *and* $\phi_Q(\cdot \mid \mathbf{R})$ *is the* Q*-dimensional normal distribution function with mean* $\mathbf{0}$ *and covariance matrix* $\mathbf{R}$. *The parameters of the model are represented by* $\boldsymbol{\theta}^\top = (\boldsymbol{\alpha}^\top, \{\text{vech}(\mathbf{R})\}^\top)$, *with* $\boldsymbol{\alpha}_{L \times 1} = (\alpha_q^{\ell_q}, \ell_q = 1, \cdots, L_q; q = 1, \cdots, Q)$, $L = \sum_{q=1}^{Q} L_q$, *and* $\text{vech}(\mathbf{R})$ *is the* $Q(Q-1)/2 \times 1$ *vector containing the unique elements of* $\mathbf{R}$.

Definition 3.1 is due to Anderson and Pemberton (1985). Further remarks concerning the model are given below.

Remark 3.2.1 *Suppose* $\text{E}(Y_q^*) = \mu_q^*$, $\text{var}(Y_q^*) = \sigma_{qq} = \sigma_q^2$, *and* $\text{cov}(Y_q^*, Y_{q'}^*) = \sigma_{qq'}$. *Then* $\mathbf{D}^{-1}(\mathbf{y}^* - \boldsymbol{\mu}^*) \sim \mathcal{N}_Q(\mathbf{0}, \mathbf{R})$, *where* $\mathbf{D} = \text{diag}(\sigma_1^2, \cdots, \sigma_Q^2)$ *and* $\boldsymbol{\mu}^* = (\mu_1^*, \cdots, \mu_Q^*)^\top$. *Hence, without loss of generality, it can be assumed that* $\boldsymbol{\mu}^* = \mathbf{0}$ *and* $\sigma_q^2 = 1 \ \forall q$.

Remark 3.2.2 *The cutpoints* $\alpha_q^{\ell_q}$ *account for the ordinal information in the data while the polychoric correlations* $r_{qq'}$ *represent the associations between the ordinal variables, for* $\ell_q = 1, \cdots, L_q + 1; q = 1, \cdots, Q$.

Remark 3.2.3 *There are a total of* $P = Q(Q-1)/2 + L$ *parameters in the model.*

Maximum likelihood estimation for the grouped continuous model is reviewed in the next section.

3.3 Maximum Likelihood Estimation

Let $\mathbf{z}_1, \cdots, \mathbf{z}_N$ be a random sample from the grouped continuous model with parameter $\boldsymbol{\theta}$, and let the number of observations such that $\mathbf{z}_i = (\ell_1, \cdots, \ell_Q)^\top$

be denoted by $n_{\ell_1 \cdots \ell_Q}$. The likelihood function $\mathcal{L}$ of the parameter $\boldsymbol{\theta}$ is then

$$\mathcal{L} = \prod_{\ell_1=1}^{L_1+1} \cdots \prod_{\ell_Q=1}^{L_Q+1} \left[\sum_{\epsilon_1=0}^{1} \cdots \sum_{\epsilon_Q=0}^{1} (-1)^{\sum_{q=1}^{Q} \epsilon_q + Q} \Phi_Q \left(\cdots, \alpha_q^{\ell_q - \epsilon_q}, \cdots \mid \mathbf{R} \right) \right]^{n_{\ell_1 \cdots \ell_Q}},$$

where $\Phi_Q(\cdot \mid \mathbf{R})$ is the Q-dimensional normal distribution function with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$. It is clear that the above expression holds in the case $Q = 2$, and it can be seen to hold for general Q by induction. First, assume it is true for $Q - 1$. Then by standard results on multivariate normal distributions (Anderson, 1984, p. 35), (3.2) can be shown to be

$$\begin{aligned} \Pr(\mathbf{z} = \boldsymbol{\ell}) &= \sum_{\epsilon_1=0}^{1} \cdots \sum_{\epsilon_{Q-1}=0}^{1} (-1)^{\sum_{q=1}^{Q-1} \epsilon_q + (Q-1)} \\ &\quad \times \int_{\alpha_Q^{\ell_Q - 1}}^{\alpha_Q^{\ell_Q}} \Phi_{Q-1} \left(\cdots, \frac{\alpha_q^{\ell_q - \epsilon_q} - r_{Qq} v_Q}{\sqrt{1 - r_{Qq}^2}}, \cdots \mid v_Q, \mathbf{R}_{\cdot Q} \right) \phi(v_Q) \mathrm{d}v_Q \\ &= \sum_{\epsilon_1=0}^{1} \cdots \sum_{\epsilon_Q=0}^{1} (-1)^{\sum_{q=1}^{Q} \epsilon_q + Q} \Phi_Q \left(\cdots, \alpha_q^{\ell_q - \epsilon_q}, \cdots \mid \mathbf{R} \right), \end{aligned}$$

where r_{Qq} is the correlation between Y_Q^* and Y_q^*, $q = 1, \cdots, Q-1$, and $\mathbf{R}_{\cdot Q}$ is the partial correlation matrix given Y_Q^*. This shows that the expression is true for Q.

The likelihood (or log-likelihood) function above is maximized to obtain the MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. As this involves evaluation of high dimensional normal integrals, which can be computationally demanding in practice, most early work on maximum likelihood estimation for the grouped continuous model has been limited to the case $Q = 2$ or 3 (e.g., Lee, 1985; Olsson, 1979). Lee and Lau (1986) looked into a generalized least squares method for the bivariate case and compared it with maximum likelihood and minimum chi-squares methods. See also Lee *et al.* (1989).

Anderson and Pemberton (1985) proposed a two-step approach to estimate $\boldsymbol{\theta}$ that relies on first marginally estimating the cutpoints, and then estimating the polychoric correlations from the bivariate likelihoods with the cutpoints replaced by their estimates. Because the second step involves maximizing the likelihood with respect only to the correlations $r_{qq'}$, it is somewhat less computationally demanding than maximum likelihood estimation.

In the next section, another alternative method, which is more conceptually appealing than the PML methods, is developed. The new method is based on the *pairwise likelihood approach* (Kuk and Nott, 2000).

3.4 Maximum Pairwise Likelihood Estimation

In view of the computational inconvenience arising from the use of the full likelihood, a natural alternative approach, motivated by the current interest in estimating equations, is to work with *pseudo-likelihoods*. The approach adopted in this section employs pairwise likelihoods in constructing a pseudo-likelihood function from which an estimating function is constructed.

Specifically, consider a pair of ordinal variables Z_q and $Z_{q'}$. The log-likelihood for $(Z_q, Z_{q'})^\top$ is

$$\begin{aligned}
\log \Pr(Z_q = \ell_q, Z_{q'} = \ell_{q'}) &= \log \left[\int_{\alpha_q^{\ell_q - 1}}^{\alpha_q^{\ell_q}} \int_{\alpha_{q'}^{\ell_{q'} - 1}}^{\alpha_{q'}^{\ell_{q'}}} \phi_2(v_q, v_{q'} \mid r_{qq'}) \mathrm{d}v_{q'} \mathrm{d}v_q \right] \\
&= \log \left[\Phi_2(\alpha_q^{\ell_q}, \alpha_{q'}^{\ell_{q'}} \mid r_{qq'}) - \Phi_2(\alpha_q^{\ell_q}, \alpha_{q'}^{\ell_{q'}-1} \mid r_{qq'}) \right. \\
&\quad \left. - \Phi_2(\alpha_q^{\ell_q - 1}, \alpha_{q'}^{\ell_{q'}} \mid r_{qq'}) + \Phi_2(\alpha_q^{\ell_q - 1}, \alpha_{q'}^{\ell_{q'}-1} \mid r_{qq'}) \right] \\
&\equiv \ell_{qq'}^{p},
\end{aligned}$$

where $\phi_2(\cdot,\cdot \mid r_{qq'})$ is the standard bivariate normal density with correlation $r_{qq'}$ and $\Phi_2(\cdot,\cdot \mid r_{qq'})$ is the corresponding distribution function. Denoting by $\ell^p_{iqq'} = \log \Pr(Z_{iq} = \ell_q, Z_{iq'} = \ell_{q'})$ for the ith observation, an overall *pairwise log-likelihood function* for $\boldsymbol{\theta}$ may then be constructed as

$$\begin{aligned} \ell^p(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \sum_{q<q'} \ell^p_{iqq'} \\ &= \sum_{q<q'} \sum_{\ell_q=1}^{L_q+1} \sum_{\ell_{q'}=1}^{L_{q'}+1} n_{\ell_q \ell_{q'}} \ell^p_{qq'}, \end{aligned} \tag{3.3}$$

where $n_{\ell_q \ell_{q'}}$ is the number of observations such that $Z_q = \ell_q, Z_{q'} = \ell_{q'}$. Note that (3.3) arises from factorizing the joint density $[Z_1, \cdots, Z_Q; \boldsymbol{\theta}]$ of $Z_1, \cdots, Z_Q$ into

$$[Z_1, Z_2; \boldsymbol{\theta}] \cdots [Z_{Q-1}, Z_Q; \boldsymbol{\theta}] = \prod_{q<q'} [Z_q, Z_{q'}; \boldsymbol{\theta}], \tag{3.4}$$

where $[Z_q, Z_{q'}]$ is the bivariate normal joint density of Z_q and $Z_{q'}$. Since (3.4) is not a proper joint density, (3.3) is not a proper log-likelihood function.

Analogous to maximum likelihood estimation, the *pairwise score vector* $\mathbf{s}_{PL}(\boldsymbol{\theta}) = \partial \ell^p(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ can be similarly defined. For the pairwise log-likelihood in (3.3), $\mathbf{s}_{PL}(\boldsymbol{\theta})$ has elements

$$\frac{\partial \ell^p(\boldsymbol{\theta})}{\partial \alpha_q^{\ell_q}} = \sum_{q',q'<q} \sum_{\ell_{q'}=1}^{L_{q'}+1} \frac{n_{\ell_q \ell_{q'}} \phi_q^{\ell_q}}{B^{\ell_{q-1}^q, \ell_{q'-1}^{q'}}} \left[\Phi(\overline{\alpha}_{q'q}^{\ell_{q'}, \ell_q}) - \Phi(\overline{\alpha}_{q'q}^{\ell_{q'}-1, \ell_q}) \right],$$

for $\ell_q = 1, \cdots, L_q + 1; q = 1, \cdots, Q$, and

$$\frac{\partial \ell^p(\boldsymbol{\theta})}{\partial r_{qq'}} = \sum_{\ell_q=1}^{L_q+1} \sum_{\ell_{q'}=1}^{L_{q'}+1} \frac{n_{\ell_q \ell_{q'}}}{B^{\ell_{q-1}^q, \ell_{q'-1}^{q'}}} \left[\phi_{qq'}^{\ell_q, \ell_{q'}} - \phi_{qq'}^{\ell_q, \ell_{q'}-1} - \phi_{qq'}^{\ell_q-1, \ell_{q'}} + \phi_{qq'}^{\ell_q-1, \ell_{q'}-1} \right],$$

for $q < q'$, where $\overline{\alpha}_{q'q}^{\ell_{q'}, \ell_q} = (\alpha_{q'}^{\ell_{q'}} - r_{qq'} \alpha_q^{\ell_q}) / \sqrt{1 - r_{qq'}^2}$, $B^{\ell_{q-1}^q, \ell_{q'-1}^{q'}} = \Phi_{qq'}^{\ell_q, \ell_{q'}} - \Phi_{qq'}^{\ell_q, \ell_{q'}-1} - \Phi_{qq'}^{\ell_q-1, \ell_{q'}} + \Phi_{qq'}^{\ell_q-1, \ell_{q'}-1}$, $\Phi_{qq'}^{\ell_q, \ell_{q'}}$ and $\phi_{qq'}^{\ell_q, \ell_{q'}}$ are the distribution function

and density, respectively, of the standard bivariate normal with correlation $r_{qq'}$, evaluated at $(\alpha_q^{\ell_q}, \alpha_{q'}^{\ell_{q'}})$ (see Plackett, 1954).

The *maximum pairwise likelihood* (MPL) estimate $\widehat{\boldsymbol{\theta}}^{PL}$ of $\boldsymbol{\theta}$ is defined as the maximizer of $\ell^p(\boldsymbol{\theta})$. It can be obtained by solving the *pairwise score equation* $\mathbf{s}_{PL}(\boldsymbol{\theta}) = \mathbf{0}$ via a modified Fisher scoring algorithm (Kuk and Nott, 2000) as follows:

$$\widehat{\boldsymbol{\theta}}^{PL,(t+1)} = \widehat{\boldsymbol{\theta}}^{PL,(t)} + \left[\sum_{q<q'} \sum_{\ell_q=1}^{L_q+1} \sum_{\ell_{q'}=1}^{L_{q'}+1} n_{\ell_q \ell_{q'}} \left(\frac{\partial \ell^p_{qq'}}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \ell^p_{qq'}}{\partial \boldsymbol{\theta}} \right)^{\top} \right]^{-1} \times \mathbf{s}_{PL}(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{PL,(t)}},$$

where $\widehat{\boldsymbol{\theta}}^{PL,(t)}$ is the MPL estimate at iteration $t = 0, 1, \cdots$. The following are suggested for the initial estimates: $\widehat{\alpha}_q^{\ell_q,(0)} = \Phi^{-1}(\sum_{\ell'_q=1}^{\ell_q} n_{\ell'_q}/N)$ and $\widehat{r}_{qq'}^{(0)} = \widetilde{r}_{qq'}$ is the sample correlation coefficient of Z_q and $Z_{q'}$, with $n_{\ell'_q}$ the number of observations such that $Z_q = \ell'_q$. Boos (1992) also refers to $\mathbf{s}_{PL}(\boldsymbol{\theta})$ as a *generalized score vector.*

Remark 3.4.1 *The pairwise likelihood approach to the grouped continuous model is attractive because it allows for a log-likelihood involving high dimensional normal integrals to be approximated by a sum of bivariate normal integrals, which can be easily evaluated. Although the pairwise likelihood approach is very similar to Poon* et al.*'s (1990) PML and Bedrick* et al.*'s (2000) pairwise PML, it is more conceptually appealing than the partition method because it entails maximizing a single objective function, the pairwise log-likelihood function, to obtain a single set of parameter estimates. Hence, there is no need to average several estimates as is done in the partition methods.*

Remark 3.4.2 *Since the pairwise log-likelihoods $\ell^p_{iqq'}$ are proper log-likelihoods, it follows that* $\mathrm{E}(-\partial^2 \ell^p_{iqq'}/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top) = \mathrm{E}(\partial\ell^p_{iqq'}/\partial\boldsymbol{\theta} \cdot \partial\ell^p_{iqq'}/\partial\boldsymbol{\theta}^\top)$*, so that there is no need to calculate second-order derivatives of $\ell^p_{iqq'}$; hence, estimation of standard errors is simplified.*

Remark 3.4.3 *Note that the pairwise likelihood approach reduces to maximum likelihood estimation in the case $Q = 2$.*

It should be noted that the pairwise likelihood approach derives from Lindsay's (1988) *composite likelihood* approach, who suggested simply pooling marginal (univariate, bivariate or otherwise) or conditional log-likelihoods additively in situations where the full likelihood is either computationally impractical to evaluate or too complicated to construct. Liang and Zeger's (1986) GEE based on an independent working covariance matrix is a notable special case. Recent examples of its applications in practice are discussed by Parner (2001), Kuk and Nott (2000), Nott and Rydén (1999), and Heagerty and Lele (1998).

3.5 Asymptotic Results

Theorem 3.1 below proves the consistency and asymptotic normality of the MPL estimator $\widehat{\boldsymbol{\theta}}^{PL}$ in § 3.4, using standard results on generalized score equations in Boos (1992) and under the following *regularity conditions* given by Heagerty and Lele (1998), based on those provided by Guyon (1995) and Crowder (1986). In what follows, let $\partial S(\boldsymbol{\theta}, \varepsilon)$ denote the boundary of a sphere of radius $\varepsilon > 0$ centered at $\boldsymbol{\theta}$.

A1 $\mathbf{s}_{PL}(\boldsymbol{\theta})$ is continuous.

A2 $\inf_{\partial S(\boldsymbol{\theta}_0,\varepsilon)}(\boldsymbol{\theta}_0-\boldsymbol{\theta})^\top \mathrm{E}_{\boldsymbol{\theta}_0}[\mathbf{s}_{PL}(\boldsymbol{\theta})] \geq \delta$ for some $\delta > 0$ and N sufficiently large.

A3 $\sup_{\partial S(\boldsymbol{\theta}_0,\varepsilon)} \|\mathbf{s}_{PL}(\boldsymbol{\theta}) - \mathrm{E}_{\boldsymbol{\theta}_0}[\mathbf{s}_{PL}(\boldsymbol{\theta})]\| \to 0$.

A4 There exists an open neighborhood $N_{\boldsymbol{\theta}_0}$ of $\boldsymbol{\theta}_0 \in \Re^P$ over which $\mathbf{s}_{PL}$ is continuously differentiable, and there exists an integrable random variable h such that for all elements of $\partial \mathbf{s}_{PL}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and all $\boldsymbol{\kappa} \in N_{\boldsymbol{\theta}_0}$, $|\, \partial \mathbf{s}_{PL}(\boldsymbol{\kappa}, \mathbf{w})/\partial\boldsymbol{\theta} \,| < h(\mathbf{w})$.

A5 There exists a limiting covariance matrix $\mathbf{V}_\infty^{(2)}$ such that $\mathbf{V}_N^{(2)} = N \times \mathrm{E}[\mathbf{s}_{PL}(\boldsymbol{\theta})\mathbf{s}_{PL}^\top(\boldsymbol{\theta})]$, where

(*i*) $\mathbf{V}_\infty^{(2)} > \mathbf{0}$ and $\mathbf{V}_N^{(2)} \geq \mathbf{V}_\infty^{(2)}$ for $N \geq m$ for some m, and

(*ii*) $\sqrt{N}\left(\mathbf{V}_N^{(2)}\right)^{-1} \mathbf{s}_{PL} \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{I})$.

A6 There exists a sequence $\mathbf{V}_N^{(1)} = \sum_{i=1}^N \sum_{q<q'} \mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}} \ell_{iqq'}^P(\boldsymbol{\theta})\right]$ of non-stochastic matrices such that

(*i*) there exists $\mathbf{V}_\infty^{(1)}$ such that $\mathbf{V}_N^{(1)} > \mathbf{V}_\infty^{(1)}$ for $N \geq m$ for some m, and

(*ii*) $\lim_{N\to\infty}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{s}_{PL}(\boldsymbol{\theta}) - \mathbf{V}_N^{(1)}\right) = \mathbf{0}$ in probability.

Conditions A1 – A3 pertain to consistency and A4 – A6 to asymptotic normality. A discussion of them is found in Heagerty and Lele (1998).

Theorem 3.1 *Under regularity conditions* A1 – A6, *the* MPL *estimator* $\widehat{\boldsymbol{\theta}}^{PL}$ *of* $\boldsymbol{\theta}$ *is consistent and satisfies*

$$\widehat{\boldsymbol{\theta}}^{PL} - \boldsymbol{\theta} \xrightarrow{\mathcal{L}} \mathcal{N}_P\left(\mathbf{0}, \mathbf{J}_P^{-1}\mathbf{K}_P\mathbf{J}_P^{-1}\right),$$

as $N \to \infty$, *where* $\mathbf{J}_P \equiv \mathbf{J}_P(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[-\partial \mathbf{s}_{PL}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^{\top}]$ *and*

$$\begin{aligned} \mathbf{K}_P &\equiv \mathbf{K}_P(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \mathrm{E}_{\boldsymbol{\theta}} \left[\left(\sum_{q<q'} \frac{\partial \ell^{p}_{iqq'}}{\partial \boldsymbol{\theta}} \right) \left(\sum_{q<q'} \frac{\partial \ell^{p}_{iqq'}}{\partial \boldsymbol{\theta}} \right)^{\top} \right]. \end{aligned}$$

Proof. First, consistency of $\widehat{\boldsymbol{\theta}}^{PL}$ follows from results of Crowder (1986), assuming conditions A1 – A3 hold.

Next, using results given in Boos (1992), note that $\mathrm{E}_{\boldsymbol{\theta}}(\mathbf{s}_{PL}) = \mathbf{0}$, with $\mathbf{s}_{PL} \equiv \mathbf{s}_{PL}(\boldsymbol{\theta})$, so that

$$\begin{aligned} \mathrm{cov}(\mathbf{s}_{PL}) &= \mathrm{E}_{\boldsymbol{\theta}} \left(\mathbf{s}_{PL} \mathbf{s}_{PL}^{\top} \right) \\ &= \mathbf{K}_P. \end{aligned}$$

From Boos (1992) and assuming conditions A4 – A6 hold, it can be shown that

$$\widehat{\boldsymbol{\theta}}^{PL} - \boldsymbol{\theta} = \mathbf{J}_P^{-1} \mathbf{s}_{PL} + o_p(N),$$

where $o_p(N) \to 0$ as $N \to \infty$. The result is now immediate.

□

Theorem 3.1 may be used to construct large-sample tests of hypotheses concerning $\boldsymbol{\theta}$. For example, a common inferential question concerns whether or not the polychoric correlations $r_{qq'}$, $q < q'$, are all equal. That is, it is of interest to test

$$H : r_{12} = r_{13} = \cdots = r_{Q-1,Q} \qquad \text{against} \qquad K : \textit{at least 1 inequality}.$$

Hypothesis H is also known as the hypothesis of *uniform polychoric correlation structure*. Note that H is equivalent to $H' : \mathbf{C}\boldsymbol{\theta} = \mathbf{0}$, where $\mathbf{C}_{Q' \times P} = (\mathbf{0}, \mathbf{I}_{Q'}, -\mathbf{1}_{Q'})$, with $\mathbf{1}_{Q'}$ the $Q' \times 1$ vector of ones and $Q' = Q(Q-1)/2 - 1$. Because $\text{rank}(\mathbf{C}) = Q'$, it follows by Theorem 3.1 that the Wald-type statistic

$$X_W^2 = (\mathbf{C}\widehat{\boldsymbol{\theta}}^{PL})^\top \left(\mathbf{C}\widehat{\mathbf{V}}^{PL}\mathbf{C}^\top\right)^{-1} (\mathbf{C}\widehat{\boldsymbol{\theta}}^{PL})$$

is asymptotically $\chi^2_{Q'}$, under H, where $\widehat{\mathbf{V}}^{PL} = (\widehat{\mathbf{J}}_P^{PL})^{-1}\widehat{\mathbf{K}}_P^{PL}(\widehat{\mathbf{J}}_P^{PL})^{-1}$, with

$$\widehat{\mathbf{K}}_P^{PL} = \sum_{i=1}^{N} \left(\sum_{q<q'} \frac{\partial \ell^p_{iqq'}}{\partial \boldsymbol{\theta}}\right)\left(\sum_{q<q'} \frac{\partial \ell^p_{iqq'}}{\partial \boldsymbol{\theta}}\right)^\top \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{PL}}, \tag{3.5}$$

and $\widehat{\mathbf{J}}_P^{PL} = \sum_{i=1}^{N}\sum_{q<q'} \left(\partial \ell^p_{iqq'}/\partial\boldsymbol{\theta}\right)\left(\partial \ell^p_{iqq'}/\partial\boldsymbol{\theta}\right)^\top$, the observed pairwise information matrix evaluated at $\widehat{\boldsymbol{\theta}}^{PL}$. The test rejects H if and only if $X_W^2 > \chi^2_{Q',\alpha}$, the $(1-\alpha)$th percentile of $\chi^2_{Q'}$.

The finite-sample performance of the MPL estimators is investigated in the next section.

3.6 Simulation Study

To assess the performance of the MPL estimators, a series of simulation experiments were conducted using the grouped continuous model with $Q = 3$. Random samples were generated from a 3-dimensional multivariate normal latent distribution with correlation matrix $\mathbf{R}^*$, and the data $\mathbf{y}_1^*, \cdots, \mathbf{y}_N^*$ were then transformed into $\mathbf{z}_1, \cdots, \mathbf{z}_N$ with the following sets of pre-assigned thresholds:

$$\text{(I)} \quad \alpha_1^1 = 0; \alpha_2^1 = -0.4, \alpha_2^2 = 0.4;$$

$$\alpha_3^1 = -0.6, \alpha_3^2 = 0, \alpha_3^3 = 0.6,$$

(II) $\alpha_1^1 = 0.5; \alpha_2^1 = -0.75, \alpha_2^2 = 0.1;$

$$\alpha_3^1 = -0.25, \alpha_3^2 = 0.3, \alpha_3^3 = 1.$$

The following correlation matrices $\mathbf{R}_{\mathrm{I}}^*$ and $\mathbf{R}_{\mathrm{II}}^*$ for cases (I) and (II), respectively, are assumed:

$$\mathbf{R}_{\mathrm{I}}^* = \begin{pmatrix} 1 & 0.5 & 0.5 \\ & 1 & 0.5 \\ & & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{R}_{\mathrm{II}}^* = \begin{pmatrix} 1 & 0.8 & 0.3 \\ & 1 & 0.4 \\ & & 1 \end{pmatrix}.$$

The correlation matrix $\mathbf{R}_{\mathrm{I}}^*$ is called the *uniform polychoric correlation matrix*. Note also that case (I) corresponds to symmetric marginal distributions for Z_1, Z_2, and Z_3, while case (II) implies that their marginal distributions are skewed. These cases are similar to those considered by Poon and Lee (1987).

For each case, samples of sizes $N = 50$ and 100 were generated and the thresholds $\alpha_1^1, \alpha_2^1, \alpha_2^2, \alpha_3^1, \alpha_3^2$, and α_3^3, and the polychoric correlations r_{12}, r_{13}, and r_{23}, were estimated using the maximum pairwise likelihood method outlined in § 3.4. This was replicated a total of $R = 50$ times, and the mean of the MPL estimates calculated. As a measure of the accuracy of the estimates, the approach of Poon and Lee (1987) and Lee and Poon (1986) was adopted and the *root mean-squared error*

$$\mathrm{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\widehat{\boldsymbol{\theta}}_r^{PL} - \boldsymbol{\theta})^2}$$

was calculated, where $\boldsymbol{\theta}$ is as defined in Definition 3.1 and $\widehat{\boldsymbol{\theta}}_r^{PL}$ is the MPL estimate of $\boldsymbol{\theta}$ for the rth replicate, $r = 1, \cdots, R$.

Tables 3.1 and 3.2 report the simulation results for case (I). Those for case (II) are displayed in Tables 3.3 and 3.4. These results indicate that the

pairwise likelihood approach can estimate both polychoric correlations and threshold parameters quite well. Specifically, the following observations may be made.

(1) The MPL estimates yielded generally small bias, which decreased with increasing sample size. In addition, the bias of the MPL estimates in case (I) was, in general, smaller than the bias of those in case (II). This suggests that MPL estimation for grouped continuous model performs better when the threshold model results in symmetric, rather than skewed, distributions for the ordinal variables.

(2) The RMSEs of the MPL estimates were generally small. As expected, increasing the sample size decreased the RMSE. From Tables 3.3 and 3.4, it appears that RMSE is smaller for large than for small polychoric correlations, which was similarly noted by Lee and Poon (1986). Furthermore, MPL estimates for case (I), which gives a symmetric distribution for the ordinal vector, yielded generally smaller RMSEs than those for the skewed ordinal distribution in case (II).

(3) Although the MPL estimates of the polychoric correlations yielded small bias, the bias were generally positive. This implies that the pairwise likelihood approach for the grouped continuous model tends to underestimate the polychoric correlations. This confirms a similar observation made by Heagerty and Lele (1998) regarding composite likelihood estimation.

The above observations are in general agreement with those made by Poon and Lee (1987) and Lee and Poon (1986) in connection with the PML method, and by Heagerty and Lele (1998) concerning the composite likelihood approach. The latter is expected since the pairwise likelihood method is a special case of composite likelihood estimation.

3.7 Conditional Grouped Continuous Model: Extension to Mixed Ordinal and Continuous Data

Consider a vector $\mathbf{y} = (Y_1, \cdots, Y_C)^\top$ of continuous variables in addition to $\mathbf{z}$. As in the grouped continuous model, a latent vector $\mathbf{y}^* \sim \mathcal{N}_Q(\mathbf{0}, \mathbf{R}^*)$ is assumed for $\mathbf{z}$, such that $\mathbf{y}$ and $\mathbf{y}^*$ are jointly normally distributed with $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}$, $\mathrm{var}(\mathbf{y}) = \boldsymbol{\Sigma}$, and $\mathrm{cov}(\mathbf{y}, \mathbf{y}^*) = \boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}^*}$. If $\boldsymbol{\Sigma}$ is the correlation matrix of $\mathbf{y}$, $\boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}^*}$ becomes the matrix containing the *polyserial* correlations of $\mathbf{y}$ and $\mathbf{y}^*$.

Conditional on $\mathbf{y}$, $\mathbf{y}^*$ is multivariate normal with mean $\boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}^*}^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ and covariance matrix

$$\begin{aligned}
\mathbf{R}^* - \boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}^*}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}^*} &= \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_Q \end{pmatrix} \begin{pmatrix} 1 & r_{12} & \cdots & r_{1Q} \\ r_{21} & 1 & \cdots & r_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ r_{Q1} & r_{Q2} & \cdots & 1 \end{pmatrix} \\
&\quad \times \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_Q \end{pmatrix} \qquad (3.6) \\
&\equiv \mathbf{DRD},
\end{aligned}$$

where $\mathbf{D}$ is the diagonal matrix of conditional standard deviations and $\mathbf{R}$ is the

symmetric matrix of conditional polychoric correlations of $\mathbf{z}$. The factorization in (3.6) follows from Seber (1984, p. 10). The *conditionally standardized* latent vector

$$\mathbf{D}^{-1}\left[\mathbf{y}^{*} - \boldsymbol{\Sigma}_{\mathbf{yy}^{*}}^{\top}\boldsymbol{\Sigma}^{-1}\left(\mathbf{y} - \boldsymbol{\mu}\right)\right] = \mathbf{D}^{-1}\mathbf{y}^{*} - \mathbf{B}\left(\mathbf{y} - \boldsymbol{\mu}\right)$$

is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$, where $\mathbf{B} = \mathbf{D}^{-1}\boldsymbol{\Sigma}_{\mathbf{yy}^{*}}^{\top}\boldsymbol{\Sigma}^{-1}$. Assuming the threshold model (3.1) for $\mathbf{y}^{*}$ and $\mathbf{z}$, the cutpoints $\alpha_{q}^{\ell_q}$ are similarly standardized as $\nu_{q}^{\ell_q} = \gamma_{q}^{\ell_q} - \boldsymbol{\beta}_{q}^{\top}\mathbf{y}$, where $\gamma_{q}^{\ell_q} = \alpha_{q}^{\ell_q}/d_q$, and $\boldsymbol{\beta}_{q}^{\top}$ is the qth row of $\mathbf{B}$, $\ell_q = 1, \cdots, L_q$, with $\gamma_{q}^{0} = -\infty$ and $\gamma_{q}^{L_q+1} = +\infty$. For some $\boldsymbol{\ell} = (\ell_1, \cdots, \ell_Q)^{\top}$, the conditional distribution $[\mathbf{z} \mid \mathbf{y}]$ of $\mathbf{z}$ given $\mathbf{y}$ is then

$$[\mathbf{z} = \boldsymbol{\ell} \mid \mathbf{y}] = \int_{\mathcal{S}} \phi_Q\left(\mathbf{v} \mid \mathbf{R}\right) \mathrm{d}\mathbf{v}, \tag{3.7}$$

where $\mathcal{S} = \{(v_1, \cdots, v_Q) : \nu_{q}^{\ell_q - 1} < v_q \leq \nu_{q}^{\ell_q}, q = 1, \cdots, Q\}$. The joint density $[\mathbf{y}, \mathbf{z}]$ of $\mathbf{y}$ and $\mathbf{z}$ is then

$$[\mathbf{y}, \mathbf{z} = \boldsymbol{\ell}] = \phi(\mathbf{y} - \boldsymbol{\mu} \mid \boldsymbol{\Sigma}) \int_{\mathcal{S}} \phi_Q\left(\mathbf{v} \mid \mathbf{R}\right) \mathrm{d}\mathbf{v}.$$

It is now possible to define the *conditional grouped continuous model* as follows.

Definition 3.2 *The vectors* $\mathbf{y}$ *and* $\mathbf{z}$ *are said to be jointly distributed according to the conditional grouped continuous model if and only if* $\mathbf{y} \sim \mathcal{N}_C(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and*

$$\begin{aligned} \Pr(\mathbf{z} = \boldsymbol{\ell} \mid \mathbf{y}) &= \Pr(Z_1 = \ell_1, \cdots, Z_Q = \ell_Q \mid \mathbf{y}) \\ &= \int_{\mathcal{S}} \phi_Q\left(\mathbf{v} \mid \mathbf{R}\right) \mathrm{d}\mathbf{v}. \end{aligned}$$

The parameters of the model are represented by $\boldsymbol{\theta}^\top = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)$, *where*

$$\begin{aligned} \boldsymbol{\theta}_1^\top &= (\boldsymbol{\mu}^\top, \{\text{vech}(\boldsymbol{\Sigma})\}^\top), \\ \boldsymbol{\theta}_2^\top &= (\boldsymbol{\gamma}^\top, \{\text{vech}(\mathbf{R})\}^\top, \boldsymbol{\beta}^\top), \end{aligned}$$

where $\boldsymbol{\gamma}^\top = (\gamma_q^{\ell_q}, \ell_q = 1, \cdots, L_q; q = 1, \cdots, Q)$, $\text{vech}(\mathbf{R})$ *is the vector containing the upper diagonal elements of* $\mathbf{R}$, *and* $\boldsymbol{\beta} = \text{vec}(\mathbf{B})$ *is the vector obtained by stacking the rows of* $\mathbf{B}$.

The conditional grouped continuous model is an extension of the grouped continuous model which applies to data with a mixture of ordinal and continuous variables. Further remarks about the model are given below.

Remark 3.7.1 *The regression parameters* $\boldsymbol{\beta}_q$, $q = 1, \cdots, Q$, *represent the polyserial correlations between* $\mathbf{y}$ *and* $\mathbf{z}$. *If* $\boldsymbol{\beta}_q = \mathbf{0}$ $\forall q$, *then* $\mathbf{y}$ *and* $\mathbf{z}$ *are independent and separate analyses suffice. In this case, the conditional probit model in (3.7) reduces to the grouped continuous model.*

Remark 3.7.2 *There are a total of* $P = C + C(C-1)/2 + Q(Q-1)/2 + CQ + L$ *parameters in the model,* $C(Q+1) + C(C-1)/2$ *parameters more than those in the grouped continuous model.*

Remark 3.7.3 *The conditional grouped continuous model was originally defined by Anderson and Pemberton (1985). It was also described by Poon and Lee (1987) in the context of polychoric and polyserial correlation estimation. Special cases were earlier studied by Tate (1955; 1954) for* $C = L = Q = 1$, *Hannan and Tate (1965) for* $L = Q = 1$, *and by Lee and Poon (1986) and*

Cox (1974) for $Q = 1, L > 1$. See also the discussion in Drasgow (1986) on polychoric and polyserial correlations.

Maximum likelihood estimation for the conditional grouped continuous model is similar to that for the grouped continuous model, except for the inclusion of the regression parameters $\boldsymbol{\beta}_q$, $q = 1, \cdots, Q$. For a sample $(\mathbf{y}_i^\top, \mathbf{z}_i^\top)^\top$, $i = 1, \cdots, N$, the likelihood is given by

$$\begin{aligned} \mathcal{L} &= \phi_C^N(\mathbf{y}_1, \cdots, \mathbf{y}_N \mid \boldsymbol{\theta}_1) \prod_{\ell_1=1}^{L_1+1} \cdots \prod_{\ell_Q=1}^{L_Q+1} \\ &\quad \times \prod_{i(\boldsymbol{\ell})} \left[\sum_{\epsilon_1=0}^{1} \cdots \sum_{\epsilon_Q=0}^{1} (-1)^{\sum_{q=1}^{Q} \epsilon_q + Q} \Phi_Q \left(\cdots, \nu_{i(\boldsymbol{\ell})q}^{\ell_q - \epsilon_q}, \cdots \mid \mathbf{R} \right) \right], \end{aligned} \tag{3.8}$$

where $\phi_Q^N(\cdot \mid \boldsymbol{\theta}_1)$ is the usual multivariate normal likelihood and the index "$i(\boldsymbol{\ell})$" refers to the ith unit in the cell of the $(L_1+1) \times \cdots \times (L_Q+1)$ contingency table for which $Z_1 = \ell_1, \cdots, Z_Q = \ell_Q$.

Maximization of (3.8) is discussed by Poon and Lee (1987), who also described a more computationally efficient PML method. See also Bedrick *et al.* (2000), Lapidus (1998) and Poon *et al.* (1990) for extension and further details.

As an alternative, the maximum pairwise likelihood approach in § 3.4 is now extended to the conditional grouped continuous model. With

$$\begin{aligned} \ell^p_{i(\ell_q, \ell_{q'})} &= \log \Pr(Z_q = \ell_q, Z_{q'} = \ell_{q'} \mid \mathbf{y}) \\ &= \log \left[\Phi_2(\nu_{i(\ell_q,\ell_{q'})q}^{\ell_q}, \nu_{i(\ell_q,\ell_{q'})q'}^{\ell_{q'}} \mid r_{qq'}) - \Phi_2(\nu_{i(\ell_q,\ell_{q'})q}^{\ell_q}, \nu_{i(\ell_q,\ell_{q'})q'}^{\ell_{q'}-1} \mid r_{qq'}) \right. \\ &\quad \left. - \Phi_2(\nu_{i(\ell_q,\ell_{q'})q}^{\ell_q-1}, \nu_{i(\ell_q,\ell_{q'})q'}^{\ell_{q'}} \mid r_{qq'}) + \Phi_2(\nu_{i(\ell_q,\ell_{q'})q}^{\ell_q-1}, \nu_{i(\ell_q,\ell_{q'})q'}^{\ell_{q'}-1} \mid r_{qq'}) \right], \end{aligned}$$

and after factorizing $[\mathbf{y}, \mathbf{z}]$ as

$$[\mathbf{y};\boldsymbol{\theta}_1][Z_1, Z_2;\boldsymbol{\theta}_2]\cdots[Z_{Q-1}, Z_Q;\boldsymbol{\theta}_2] = [\mathbf{y};\boldsymbol{\theta}_1]\prod_{q<q'}[Z_q, Z_{q'};\boldsymbol{\theta}_2], \tag{3.9}$$

the pairwise log-likelihood function can then be defined as

$$\begin{aligned}\ell^p(\boldsymbol{\theta}) &= \ell_1^p(\boldsymbol{\theta}_1) + \sum_{q<q'}\sum_{\ell_q=1}^{L_q+1}\sum_{\ell_{q'}=1}^{L_{q'}+1}\sum_{i(\ell_q,\ell_{q'})}\ell^p_{i(\ell_q,\ell_{q'})},\\ &= \ell_1(\boldsymbol{\theta}_1) + \ell_2^p(\boldsymbol{\theta}_2),\end{aligned} \tag{3.10}$$

where $\ell_1(\boldsymbol{\theta}_1)$ is the usual multivariate normal log-likelihood function and the index "$i(\ell_q, \ell_{q'})$" refers to the ith unit in the cell of the $(L_q + 1) \times (L_{q'} + 1)$ contingency table for which $Z_q = \ell_q, Z_{q'} = \ell_{q'}$. Maximizing (3.10) yields the MPL estimate $\widehat{\boldsymbol{\theta}}^{PL}$. The usual MLEs $\overline{\mathbf{y}}$ and vech($\mathbf{S}$) (Mardia *et al.*, 1979, pp. 103-105) are obtained for $\boldsymbol{\theta}_1$ while that for $\boldsymbol{\theta}_2$ entails an iterative method as in § 3.4.

The elements of the pairwise score vector $\mathbf{s}_{PL}(\boldsymbol{\theta}_2) = \partial\ell_2^p(\boldsymbol{\theta}_2)/\partial\boldsymbol{\theta}_2$ are given by the following:

$$\frac{\partial\ell_2^p(\boldsymbol{\theta}_2)}{\partial\gamma_{\ell_q}^{\ell_q}} = \sum_{q',q'<q}\sum_{\ell_{q'}=1}^{L_{q'}+1}\sum_{i(\ell_q,\ell_{q'})}\frac{\phi^{\ell_q}_{i(\ell_q,\ell_{q'})q}}{B^{\ell_{q-1},\ell_{q'-1}}_{i(\ell_q,\ell_{q'})qq'}}\left[\Phi(\overline{\nu}^{\ell_{q'},\ell_q}_{i(\ell_q,\ell_{q'})qq'}) - \Phi(\overline{\nu}^{\ell_{q'-1},\ell_q}_{i(\ell_q,\ell_{q'})qq'})\right],$$

$$\begin{aligned}\frac{\partial\ell_2^p(\boldsymbol{\theta}_2)}{\partial\tau_{qq'}} = \sum_{\ell_q=1}^{L_q+1}\sum_{\ell_{q'}=1}^{L_{q'}+1}\sum_{i(\ell_q,\ell_{q'})}\frac{1}{B^{\ell_{q-1},\ell_{q'-1}}_{i(\ell_q,\ell_{q'})qq'}}&\left[\phi^{\ell_q,\ell_{q'}}_{i(\ell_q,\ell_{q'})qq'} - \phi^{\ell_q-1,\ell_{q'}}_{i(\ell_q,\ell_{q'})qq'}\right.\\ &\left.-\phi^{\ell_q,\ell_{q'}-1}_{i(\ell_q,\ell_{q'})qq'} + \phi^{\ell_q-1,\ell_{q'}-1}_{i(\ell_q,\ell_{q'})qq'}\right],\end{aligned}$$

$$\begin{aligned}\frac{\partial\ell_2^p(\boldsymbol{\theta}_2)}{\partial\boldsymbol{\beta}_q} = \sum_{q',q'<q}\sum_{\ell_q=1}^{L_q+1}\sum_{\ell_{q'}=1}^{L_{q'}+1}\sum_{i(\ell_q,\ell_{q'})}&\frac{\mathbf{y}_{i(\ell_q,\ell_{q'})}}{B^{\ell_{q-1},\ell_{q'-1}}_{i(\ell_q,\ell_{q'})qq'}}\\ &\times\left[\phi^{\ell_q}_{i(\ell_q,\ell_{q'})q}\Phi(\overline{\nu}^{\ell_{q'},\ell_q}_{i(\ell_q,\ell_{q'})q'q}) - \phi^{\ell_q-1}_{i(\ell_q,\ell_{q'})q}\Phi(\overline{\nu}^{\ell_{q'},\ell_q-1}_{i(\ell_q,\ell_{q'})q'q})\right.\\ &\left.-\phi^{\ell_q}_{i(\ell_q,\ell_{q'})q}\Phi(\overline{\nu}^{\ell_{q'}-1,\ell_q}_{i(\ell_q,\ell_{q'})q'q}) + \phi^{\ell_q-1}_{i(\ell_q,\ell_{q'})q}\Phi(\overline{\nu}^{\ell_{q'}-1,\ell_q-1}_{i(\ell_q,\ell_{q'})q'q})\right],\end{aligned}$$

where $B^{\ell_q,\ell_{q'}}_{\mathrm{i}(\ell_q,\ell_{q'})qq'} = \Phi^{\ell_q,\ell_{q'}}_{\mathrm{i}(\ell_q,\ell_{q'})qq'} - \Phi^{\ell_q-1,\ell_{q'}}_{\mathrm{i}(\ell_q,\ell_{q'})qq'} - \Phi^{\ell_q,\ell_{q'}-1}_{\mathrm{i}(\ell_q,\ell_{q'})qq'} + \Phi^{\ell_q-1,\ell_{q'}-1}_{\mathrm{i}(\ell_q,\ell_{q'})qq'}$, with $\Phi^{\ell_q,\ell_{q'}}_{\mathrm{i}(\ell_q,\ell_{q'})qq'}$ and $\phi^{\ell_q,\ell_{q'}}_{\mathrm{i}(\ell_q,\ell_{q'})qq'}$ are the distribution and density functions, respectively, of the standard bivariate normal with correlation $r_{qq'}$ evaluated at $(\nu^{\ell_q}_{\mathrm{i}(\ell_q,\ell_{q'})q}, \nu^{\ell_{q'}}_{\mathrm{i}(\ell_q,\ell_{q'})q'})$ and $\overline{\nu}^{\ell_{q'},\ell_q}_{\mathrm{i}(\ell_q,\ell_{q'})q'q} = (\nu^{\ell_{q'}}_{\mathrm{i}(\ell_q,\ell_{q'})q'} - r_{qq'}\nu^{\ell_q}_{\mathrm{i}(\ell_q,\ell_{q'})q})/\sqrt{1-r^2_{qq'}}$. To get $\widehat{\boldsymbol{\theta}}_2^{PL}$, the same modified Fisher scoring algorithm recommended for the grouped continuous model may be used. In addition to the initial estimates in § 3.4, $\widehat{\boldsymbol{\beta}}_q^{(0)} = \mathbf{S}^{-1}\widetilde{\boldsymbol{\sigma}}_q/\sqrt{1-\widetilde{\boldsymbol{\sigma}}_q^{\top}\mathbf{S}^{-1}\widetilde{\boldsymbol{\sigma}}_q}$ is suggested, where $\widetilde{\boldsymbol{\sigma}}_q$ is the sample covariance vector between $\mathbf{y}$ and Z_q, the cth element of which is given by $\widetilde{\sigma}_{qc} = \tilde{r}_{qc}\sqrt{s_{cc}}$, with s_{cc} the cth diagonal element of $\mathbf{S}$.

Theorem 3.2 *Assuming the regularity conditions* A1 – A6 *hold, the* MPL *estimator* $\widehat{\boldsymbol{\theta}}^{PL}$ *of* $\boldsymbol{\theta}$ *is consistent and satisfies*

$$\widehat{\boldsymbol{\theta}}^{PL} - \boldsymbol{\theta} \xrightarrow{\mathcal{L}} \mathcal{N}_P(\mathbf{0}, \mathbf{V}),$$

as $N \to \infty$, *where*

$$\mathbf{V} = \begin{pmatrix} \frac{1}{N}\mathcal{I}^{-1}_{P_1}(\boldsymbol{\theta}_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}^{-1}_{P_2}\mathbf{K}_{P_2}\mathbf{J}^{-1}_{P_2} \end{pmatrix},$$

$\mathcal{I}^{-1}_{P_1}(\boldsymbol{\theta}_1) = N\mathrm{diag}(\boldsymbol{\Sigma}/N, \boldsymbol{\Upsilon})$ *with* $\boldsymbol{\Upsilon}_{(P_1-C)\times(P_1-C)}$ *containing the asymptotic variances and covariances of the unique elements of* $\mathbf{S}$, $\mathbf{J}_{P_2} = \mathrm{E}_{\boldsymbol{\theta}}[-\partial \mathbf{s}_{PL}(\boldsymbol{\theta}_2)/\partial\boldsymbol{\theta}_2^{\top}]$,

$$\mathbf{K}_{P_2} = \sum_{i=1}^{N} \mathrm{E}_{\boldsymbol{\theta}}\left[\left(\sum_{q<q'} \frac{\partial \ell^{p}_{iqq'}}{\partial \boldsymbol{\theta}_2}\right)\left(\sum_{q<q'} \frac{\partial \ell^{p}_{iqq'}}{\partial \boldsymbol{\theta}_2}\right)^{\top}\right],$$

$\ell^{p}_{iqq'}$ *is the pairwise log-likelihood of* Z_q *and* $Z_{q'}$ *for the* i*th observation,* $P_1 = C + C(C-1)/2$ *and* $P_2 = P - P_1$.

Proof. The proof of consistency follows in the same way as that in Theorem 3.1. Following Boos (1992), define the *pairwise information matrix* $\mathbf{J}_P \equiv \mathbf{J}_P(\boldsymbol{\theta})$

as follows:

$$\begin{aligned}\mathbf{J}_P &= \mathrm{E}_{\boldsymbol{\theta}}\left[-\frac{\partial^2 \ell_1^P(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right] + \mathrm{E}_{\boldsymbol{\theta}}\left[-\frac{\partial^2 \ell_2^P(\boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right] \\ &= \begin{pmatrix} N\mathcal{I}_{P_1}(\boldsymbol{\theta}_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{P_2} \end{pmatrix},\end{aligned}$$

where the last equality follows from the fact that $\mathbf{J}_{P_2} = \mathrm{E}_{\boldsymbol{\theta}}[-\partial \mathbf{s}_{PL}(\boldsymbol{\theta}_2)/\partial\boldsymbol{\theta}_2^\top]$ and $\widehat{\boldsymbol{\theta}}_1^{PL} = \widehat{\boldsymbol{\theta}}_1$. Also, define $\mathbf{K}_P \equiv \mathbf{K}_P(\boldsymbol{\theta})$ as

$$\begin{aligned}\mathbf{K}_P &= \sum_{i=1}^{N} \mathrm{E}_{\boldsymbol{\theta}}\left[\left(\sum_{q<q'} \frac{\partial \ell_{iqq'}^P}{\partial \boldsymbol{\theta}_2}\right)\left(\sum_{q<q'} \frac{\partial \ell_{iqq'}^P}{\partial \boldsymbol{\theta}_2}\right)^\top\right] \\ &= \begin{pmatrix} N\mathcal{I}_{P_1}(\boldsymbol{\theta}_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{P_2} \end{pmatrix},\end{aligned}$$

where the last equality follows from the fact that $\mathbf{K}_{P_1}(\boldsymbol{\theta}_1) = N\mathcal{I}_{P_1}(\boldsymbol{\theta}_1)$ (Mardia *et al.*, 1979, p. 98), with $\ell_i^P = \sum_{q<q'} \ell_{iqq'}^P$. The elements of $\boldsymbol{\Upsilon}$ can be obtained by applying Theorem 3.4.4 of Anderson (1984, pp. 81-82), making note of the fact that $\mathbf{S}$ is uncorrected for bias.

From Theorem 3.1, it follows that since

$$\widehat{\boldsymbol{\theta}}^{PL} - \boldsymbol{\theta} = \mathbf{J}_P^{-1}\begin{pmatrix} \frac{\partial}{\partial\boldsymbol{\theta}_1}\ell_1(\boldsymbol{\theta}_1) \\ \mathbf{s}_{PL}(\boldsymbol{\theta}_2) \end{pmatrix} + o_p(N),$$

where $o_p(N) \xrightarrow{p} 0$ as $N \to \infty$, $\widehat{\boldsymbol{\theta}}^{PL} - \boldsymbol{\theta}$ is asymptotically multivariate normal with asymptotic mean $\mathbf{0}$ and asymptotic covariance matrix

$$\begin{aligned}\mathbf{V} &= \mathbf{J}_P^{-1}\mathbf{K}_P\mathbf{J}_P^{-1} \\ &= \begin{pmatrix} \frac{1}{N}\mathcal{I}_{P_1}^{-1}(\boldsymbol{\theta}_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{P_2}^{-1}\mathbf{K}_{P_2}\mathbf{J}_{P_2}^{-1} \end{pmatrix}.\end{aligned}$$

□

Note that Theorem 3.2 extends Theorem 3.1 to the conditional grouped continuous model.

Similar to § 3.5, generalized Wald-type tests can be constructed to test hypotheses concerning $\boldsymbol{\theta}$. Further discussion on hypothesis-testing is deferred until Chapter 4.

3.8 Discussion

In this chapter, an alternative method to maximum likelihood estimation was proposed for the grouped continuous model and its extension, the conditional grouped continuous model. By working with pairwise likelihoods instead of the full likelihood of the model, high dimensional numerical integration is avoided. The pairwise likelihood method is thus computationally simple, and properties such as consistency and asymptotic normality of the estimators readily follow from standard theory.

Moreover, the pairwise likelihood method provides a viable alternative to the PML methods of Poon and Lee (1987) and Bedrick *et al.* (2000). Unlike the latter, the former simultaneously estimates the parameters yielding a single set of estimates. Thus, the problem of having to deal with several estimates required in PML methods is avoided. This is accomplished by specifying a single objective function, the pairwise log-likelihood function, which is maximized to obtain the estimates. In this respect, the pairwise likelihood method is more conceptually appealing than PML methods.

Finally, maximum pairwise likelihood estimates of the parameters of the grouped continuous model appear to perform quite well, as indicated by the simulation results. Bias was minimal and the root mean-squared errors were

generally small.

Table 3.1: *Maximum Pairwise Likelihood Estimates based on 50 Random Samples of Size $N = 50$ from the Grouped Continuous Model with $Q = 3$ and Parameters given by Case* (I).

Parameter	True	Ave. MPL Estimate	Bias	Relative Bias (%)	RMSE
		Polychoric Correlations			
r_{12}	0.5	0.468	0.032	6.4	0.067
r_{13}	0.5	0.489	0.011	2.204	0.081
r_{23}	0.5	0.482	0.018	3.559	0.13
		Thresholds			
α_1^1	0	-0.052	0.052	—	0.107
α_2^1	-0.4	-0.344	-0.056	14.068	0.209
α_2^2	0.4	0.418	-0.018	-4.528	0.219
α_3^1	-0.6	-0.612	0.012	2.043	0.122
α_3^2	0	0.186	-0.186	—	0.156
α_3^3	0.6	0.591	0.009	1.433	0.182

NOTE: *Shown are the bias, relative bias, and root mean-squared error of the maximum pairwise likelihood estimates for the grouped continuous model with $Q = 3$ and polychoric correlation matrix $\mathbf{R}_1^*$. Note that relative bias=(bias/true)$\times$100.*

Table 3.2: *Maximum Pairwise Likelihood Estimates based on 50 Random Samples of Size $N = 100$ from the Grouped Continuous Model with $Q = 3$ and Parameters given by Case* (I).

Parameter	True	Ave. MPL Estimate	Bias	Relative Bias (%)	RMSE
		Polychoric Correlations			
r_{12}	0.5	0.473	0.027	5.323	0.09
r_{13}	0.5	0.495	0.005	1.031	0.137
r_{23}	0.5	0.486	0.014	2.893	0.069
		Thresholds			
α_1^1	0	-0.023	0.023	—	0.093
α_2^1	-0.4	-0.416	0.016	-4.135	0.096
α_2^2	0.4	0.397	0.003	0.69	0.123
α_3^1	-0.6	-0.59	-0.01	1.597	0.131
α_3^2	0	-0.051	0.051	—	0.099
α_3^3	0.6	0.601	-0.001	-0.022	0.098

NOTE: *Shown are the bias, relative bias, and root mean-squared error of the maximum pairwise likelihood estimates for the grouped continuous model with $Q = 3$ and polychoric correlation matrix $\mathbf{R}_1^*$. Note that relative bias=(bias/true)×100.*

Table 3.3: *Maximum Pairwise Likelihood Estimates based on 50 Random Samples of Size $N = 50$ from the Grouped Continuous Model with $Q = 3$ and Parameters given by Case* (II).

Parameter	True	Ave. MPL Estimate	Bias	Relative Bias (%)	RMSE
		Polychoric Correlations			
r_{12}	0.8	0.775	0.025	3.108	0.041
r_{13}	0.3	0.265	0.035	11.59	0.257
r_{23}	0.4	0.371	0.029	7.214	0.223
		Thresholds			
α_1^1	0.5	0.485	0.015	3.018	0.196
α_2^1	-0.75	-0.814	0.064	-8.514	0.205
α_2^2	0.1	0.075	0.025	5.185	0.193
α_3^1	-0.25	-0.238	-0.012	4.782	0.196
α_3^2	0.3	0.282	0.018	6.108	0.184
α_3^3	1	0.98	0.02	2.024	0.212

NOTE: *Shown are the bias, relative bias, and root mean-squared error of the maximum pairwise likelihood estimates for the grouped continuous model with $Q = 3$ and polychoric correlation matrix $\mathbf{R}_{II}^*$. Note that relative bias=(bias/true)×100.*

Table 3.4: *Maximum Pairwise Likelihood Estimates based on 50 Random Samples of Size $N = 100$ from the Grouped Continuous Model with $Q = 3$ and Parameters given by Case* (II).

Parameter	True	Ave. MPL Estimate	Bias	Relative Bias (%)	RMSE
		Polychoric Correlations			
r_{12}	0.8	0.793	0.007	0.854	0.022
r_{13}	0.3	0.332	-0.032	-10.537	0.1
r_{23}	0.4	0.419	-0.019	-4.869	0.024
		Thresholds			
α_1^1	0.5	0.502	-0.002	-0.365	0.078
α_2^1	-0.75	-0.776	0.026	-3.474	0.135
α_2^2	0.1	0.097	0.003	3.314	0.139
α_3^1	-0.25	-0.276	0.026	-10.212	0.063
α_3^2	0.3	0.314	-0.014	-4.703	0.084
α_3^3	1	1.049	-0.049	-4.902	0.183

NOTE: *Shown are the bias, relative bias, and root mean-squared error of the maximum pairwise likelihood estimates for the grouped continuous model with $Q = 3$ and polychoric correlation matrix $\mathbf{R}_{\text{II}}^*$. Note that relative bias=(bias/true)$\times$100.*

Chapter 4

General Mixed-Data Model: Extension of General Location and Grouped Continuous Models

4.1 Introduction

The previous chapters introduced two models that have been used in practice to analyze multivariate data consisting of mixtures of qualitative and quantitative variables. Chapter 2 introduced the general location model (e.g., Krzanowski, 1993; Olkin and Tate, 1961) for mixed binary (i.e., nominal) and continuous data while Chapter 3 studied the conditional grouped continuous model (Poon and Lee, 1987; Anderson and Pemberton, 1985), a latent variable model for mixed ordinal and continuous data. In this chapter, these models are unified into a single general model that can be used in the analysis of multivariate mixed data that include nominal, ordinal and continuous variables. The model can be viewed as extensions of the general location model to mixed data with ordinal variables in addition to binary and continuous variables and

of the conditional grouped continuous model to mixed nominal, ordinal and continuous data.

Examples of mixed multivariate data with variables measured on an ordinal scale, along with nominal and continuous outcomes, abound in the health and social sciences. Little and Schluchter (1985) present data from the St. Louis Risk Research Project. This is an observational study to assess the effects of parental psychological disorders on various aspects of child development. Variables in the study include parental risk group (nominal), high or low frequency of adverse symptoms in each child (ordinal), and the child's standardized reading and verbal test scores (continuous). The data have been analyzed in various contexts by Schafer (1997), Fitzmaurice and Laird (1997), and Little and Rubin (1987). As another example, Koepsel *et al.* (1981), also cited in Fisher and Van Bell (1993, pp. 680-683), analyzed data from 281 patients who underwent appendectomies and considered a variety of nominal, ordinal and continuous risk factors as they relate to the occurrence (or absence) of perforation of the appendix.

An obvious, but often very inefficient, approach to handling mixed data is to convert one type of variable to another, as discussed by Anderberg (1973, Chapter 3), and then to employ appropriate standard methods. Two approaches have been taken. The first transforms qualitative into quantitative variables via some scoring scheme, and then employs standard methods for the analysis of quantitative data (see, e.g., Cox and Wermuth, 1996, pp. 81-86). The second categorizes quantitative variables and the analysis then proceeds

as if the data were qualitative. Although these two approaches are simple enough and appear to work in practice (e.g., Truett *et al.*, 1967), the crude approach of coding qualitative variables in the former and categorizing quantitative variables in the latter make them conceptually unattractive, less meaningful and unsatisfactory in many applications (Krzanowski, 1993; Olsson *et al.*, 1979; Bishop *et al.*, 1975, pp. 358-361).

A model-based alternative is possible by specifying a model for the joint distribution of qualitative and quantitative variables. However, specifying a model for the joint distribution which can simultaneously deal with the different measurement levels of the variables is not straightforward. Possible models include the general location and conditional grouped continuous models, with the former treating ordinal variables as nominal variables and the latter treating nominal variables as ordinal variables (see Figure 4.1). These models, however, are inadequate, and hence, inappropriate, for two reasons. First, they fail to account for the different levels of measurement in the data—neither assigning ordinal scores to nominal variables nor treating ordinal scores without regard to their position on some scale makes full and correct use of the information contained in the data. Second, they do not provide a mechanism for explicitly incorporating correlations between nominal and ordinal variables, and thus fail to distinguish correlations between nominal and continuous variables from those between ordinal and continuous variables.

There is thus a need for a model that can be used for the case of mixed data with nominal, ordinal and continuous variables, which addresses

the shortcomings of the general location and conditional grouped continuous models. The development of such a general model is precisely the focus of this chapter.

Suppose that $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$ denote vectors of binary (i.e., nominal), continuous, and ordinal variables, respectively. The joint distribution $[\mathbf{x}, \mathbf{y}, \mathbf{z}]$ can then be factorized as follows:

$$[\mathbf{x}, \mathbf{y}, \mathbf{z}] = [\mathbf{x}][\mathbf{y} \mid \mathbf{x}][\mathbf{z} \mid \mathbf{x}, \mathbf{y}], \tag{4.1}$$

where $[\mathbf{x}], [\mathbf{y} \mid \mathbf{x}]$, and $[\mathbf{z} \mid \mathbf{x}, \mathbf{y}]$ denote, respectively, the marginal distribution of $\mathbf{x}$, the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$, and the conditional distribution of $\mathbf{z}$ given $\mathbf{x}$ and $\mathbf{y}$. The expression in (4.1) is exactly the same factorization used for the general location model, only taken one step further in view of the inclusion of the ordinal vector $\mathbf{z}$. By supposing the existence of a continuous latent vector $\mathbf{y}^*$ underlying $\mathbf{z}$, a multivariate normal distribution may be assumed for the conditional distribution $[\mathbf{y}, \mathbf{y}^* \mid \mathbf{x}]$. A threshold model as in Chapter 3 can then be postulated for $\mathbf{z}$ and its corresponding latent vector $\mathbf{y}^*$, so that $[\mathbf{z} \mid \mathbf{x}, \mathbf{y}]$ is specified through $[\mathbf{y}^* \mid \mathbf{x}, \mathbf{y}]$. Besides taking into account the various measurement levels of $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$, this approach allows for correlations between $\mathbf{x}$ and $\mathbf{y}$, between $\mathbf{x}$ and $\mathbf{z}$, and between $\mathbf{y}$ and $\mathbf{z}$ to be incorporated, implicitly and explicitly, into the model.

Similar models have previously appeared in the literature. Catalano (1997) proposed a regression model for mixed bivariate data made up of continuous and ordinal outcomes arising in developmental toxicology. His model is based on an underlying latent variable for the ordinal outcome, and assumes

a joint bivariate normal distribution for the latent variable and the continuous outcome. Sammel *et al.* (1997) also introduced a general class of latent variable models that accommodates a mixture of discrete and continuous variables from an exponential family. However, these models do not apply to the more general situation of mixed data being considered in this chapter.

Recently, Bedrick *et al.* (2000) and Lapidus (1998), building on earlier work by Poon and Lee (1987; 1986), considered the conditional grouped continuous model for mixed continuous and ordinal data from several populations to estimate the *Mahalanobis distance* (Mardia *et al.*, 1979, p. 31) between them. They proposed a modification of the *partition maximum likelihood* (PML) method (Lee and Poon, 1986) by considering pairwise (or bivariate) likelihoods in the estimation of the ordinal data parameters.

In the absence of nominal data, the model developed in this chapter reduces to the conditional grouped continuous model. Likewise, it specializes into the general location model in the case of mixed data with only binary and continuous variables. In this respect, the model may be viewed as generalizing these two models to data with mixtures of nominal, ordinal and continuous variables.

The chapter is organized as follows. The proposed model is developed in § 4.2. The model is further explored by considering a special case in § 4.2.1. Estimation of the model parameters is considered in § 4.3 and § 4.4. A full likelihood-based approach as well as an alternative based on *pairwise likelihoods* (Kuk and Nott, 2000), considered earlier in Chapter 3, are explored and

algorithms for implementing them are presented. Further statistical inference based on asymptotic results on the estimators are presented in § 4.6. Finally, the chapter concludes with a discussion in § 4.7.

4.2 General Mixed-Data Model

In this section, a general model that includes the general location and (conditional) grouped continuous models as special cases is developed. Notations due to Bedrick *et al.* (2000), Lapidus (1998), and Poon and Lee (1987) are adopted with slight modifications. The development of the model parallels those of the general location and grouped continuous models in Chapters 2 and 3.

Let $\mathbf{u}$ denote the $D \times 1$ vector of nominal variables, with the dth component of $\mathbf{u}$ having s_d possible states ($d = 1, \cdots, D$). The vector $\mathbf{u}$ then defines a contingency table with $S = \prod_{d=1}^{D} s_d$ states, one for each possible value of $\mathbf{u}$. As in Chapter 2, the index $s = 1, \ldots, S$ is used to refer to the states. From Remark 2.2.1, an $S \times 1$ vector $\mathbf{x} = (X_1, \cdots, X_S)^\top$ can be defined such that X_s is either 0 or 1 depending on whether $\mathbf{u}$ falls in state s or not ($\sum_{s=1}^{S} X_s = 1$). Following the notation in Chapter 2, the index "(s)" is used to mean that the observation falls in state s, as in $\mathbf{x}_{(s)}$, the vector $\mathbf{x}$ with $X_s = 1$.

By Definition 2.1 of the general location model, $\mathbf{x}$ is modelled by a product multinomial distribution $[\mathbf{x}; \boldsymbol{\pi}] = \prod_{s=1}^{S} \pi_s^{\mathbf{x}_{(s)}^\top \mathbf{x}}$, where $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_S)^\top$ is the vector of state probabilities ($\sum_{s=1}^{S} \pi_s = 1$).

Define the vector $(\mathbf{y}^\top, \mathbf{y}^{*\top})^\top$, where $\mathbf{y}$ is the $C \times 1$ vector of continuous

variables and $\mathbf{y}^*$ the $Q \times 1$ vector of unobservable latent variables. By the general location model, $(\mathbf{y}^\top, \mathbf{y}^{*\top})^\top$ is modelled as conditionally multivariate normal with mean $\boldsymbol{\eta}_s$ and common covariance matrix $\boldsymbol{\Gamma} > \mathbf{0}$, given $\mathbf{x} = \mathbf{x}_{(s)}$, with $\boldsymbol{\eta}_s$ and $\boldsymbol{\Gamma}$ partitioned accordingly as

$$\begin{aligned} \boldsymbol{\eta}_s &= \begin{pmatrix} \boldsymbol{\mu}_s \\ \boldsymbol{\mu}_s^* \end{pmatrix}, \\ \boldsymbol{\Gamma} &= \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}^*} \\ \boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}^*}^\top & \boldsymbol{\Sigma}^* \end{pmatrix}. \end{aligned} \tag{4.2}$$

The $CS \times 1$ stacked vector of state means of $\mathbf{y}$ is denoted as $\boldsymbol{\mu}$.

In lieu of the vector $\mathbf{y}^*$, suppose a vector of ordinal variables $\mathbf{z} = (Z_1, \cdots, Z_Q)^\top$ is observed. The latent relationship between $\mathbf{y}^*$ and $\mathbf{z}$ is defined by the threshold model in Chapter 3:

$$\begin{aligned} Z_q = a_q^1 \quad &\Longleftrightarrow \quad -\infty < Y_q^* \leq \alpha_q^1, \\ Z_q = a_q^{\ell_q} \quad &\Longleftrightarrow \quad \alpha_q^{\ell_q - 1} < Y_q^* \leq \alpha_q^{\ell_q}, \quad (\ell_q = 2, \cdots, L_q) \\ Z_q = a_q^{L_q+1} \quad &\Longleftrightarrow \quad \alpha_q^{L_q} < Y_q^* < +\infty, \end{aligned} \tag{4.3}$$

where Y_q^* is the qth element of $\mathbf{y}^*$, $\{\alpha_q^0 = -\infty, \alpha_q^1, \cdots, \alpha_q^{L_q}, \alpha_q^{L_q+1} = +\infty\}$ are the unknown cutpoints, and $a_q^1 < a_q^2 < \cdots < a_q^{L_q+1}$ are the ordinal scores for Z_q, $q = 1, \cdots, Q$. Note that the set of thresholds as well as the scores may vary for each ordinal variable in $\mathbf{z}$ but is constant across states. As in Chapter 3, it is assumed that $a_q^{\ell_q} = \ell_q$, $\ell_q = 1, \cdots, L_q + 1$.

Now suppose $\mathbf{x} = \mathbf{x}_{(s)}$. Then, under the general location model, the conditional distribution of $\mathbf{y}^*$, given $\mathbf{x}_{(s)}$ and $\mathbf{y}$, is multivariate normal with

mean $\boldsymbol{\mu}_s^* + \boldsymbol{\Sigma}_{\mathbf{yy}^*}^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}_s)$ and covariance matrix

$$\begin{aligned}\boldsymbol{\Sigma}^* - \boldsymbol{\Sigma}_{\mathbf{yy}^*}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\mathbf{yy}^*} &= \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_Q \end{pmatrix} \begin{pmatrix} 1 & r_{12} & \cdots & r_{1Q} \\ r_{21} & 1 & \cdots & r_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ r_{Q1} & r_{Q2} & \cdots & 1 \end{pmatrix} \\ &\quad \times \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_Q \end{pmatrix} \qquad (4.4) \\ &\equiv \mathbf{DRD},\end{aligned}$$

where $\mathbf{D}$ is the diagonal matrix of conditional standard deviations and $\mathbf{R}$ is the symmetric matrix of conditional correlations of $\mathbf{z}$ (referred to earlier as *polychoric correlations*), given $\mathbf{x}_{(s)}$ and $\mathbf{y}$.

Note that the factorization (4.4) is the usual factorization of the dispersion matrix in terms of the correlation matrix (Seber, 1984, p. 10), as was previously done in Chapter 3. To avoid overparameterizing the model, state S is fixed as a *reference state* and $\boldsymbol{\mu}_s$ and $\boldsymbol{\mu}_s^*$ $(s \neq S)$ are defined as $\boldsymbol{\mu}_s = \boldsymbol{\xi} + \boldsymbol{\xi}_s$ and $\boldsymbol{\mu}_s^* = \boldsymbol{\xi}^* + \boldsymbol{\xi}_s^*$, where $\boldsymbol{\mu}_S = \boldsymbol{\xi}$ and $\boldsymbol{\mu}_S^* = \boldsymbol{\xi}^*$, the means of $\mathbf{y}$ and $\mathbf{y}^*$, respectively, for state S, and $\boldsymbol{\xi}_s$ and $\boldsymbol{\xi}_s^*$ are the effects of state $s = 1, \cdots, S-1$, relative to that of state S.

Similar to Chapter 3, it follows that

$$\mathbf{D}^{-1}\left[\mathbf{y}^* - \boldsymbol{\mu}_s^* - \boldsymbol{\Sigma}_{\mathbf{yy}^*}^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu}_s)\right] = \begin{cases} \mathbf{D}^{-1}(\mathbf{y}^* - \boldsymbol{\xi}^*) - \boldsymbol{\tau}_s - \mathbf{B}(\mathbf{y} - \boldsymbol{\xi}) & s \neq S \\ \mathbf{D}^{-1}(\mathbf{y}^* - \boldsymbol{\xi}^*) - \mathbf{B}(\mathbf{y} - \boldsymbol{\xi}) & s = S \end{cases}$$

is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$, given $\mathbf{x}_{(s)}$ and $\mathbf{y}$. Here, $\boldsymbol{\tau}_s = \mathbf{D}^{-1}\boldsymbol{\xi}_s^* - \mathbf{B}\boldsymbol{\xi}_s$ and $\mathbf{B} = \mathbf{D}^{-1}\boldsymbol{\Sigma}_{\mathbf{yy}^*}^\top \boldsymbol{\Sigma}^{-1}$. From the above parameterization and similar to the usual development of latent variable models

(Wiley, 1973), it may be assumed without loss of generality that $\mathbf{\Sigma}^* = \mathbf{R}^*$, the correlation matrix of $\mathbf{y}^*$. Note that $\boldsymbol{\mu}_s^*$ was not assumed to be $\mathbf{0}$ so that the effect of $\mathbf{x}$ does not vanish with that for $\mathbf{y}$.

Similarly, the (conditional) cutpoints $\{\alpha_q^1, \cdots, \alpha_q^{L_q}\}$, $q = 1, \cdots, Q$, may be *standardized* as $\gamma_q^{\ell_q} - \tau_{sq} - \boldsymbol{\beta}_q^\top \mathbf{y}$, where $\gamma_q^{\ell_q} = \alpha_q^{\ell_q}/d_q - (\xi_q^*/d_q - \boldsymbol{\beta}_q^\top \boldsymbol{\xi})$, ξ_q^* is the qth element of $\boldsymbol{\xi}^*$, τ_{sq} is the qth element of $\boldsymbol{\tau}_s$, and $\boldsymbol{\beta}_q^\top$ is the qth row of $\mathbf{B}$, $\ell_q = 1, \cdots, L_q$. Here, $\tau_{Sq} = 0\ \forall q$, $\gamma_q^0 = -\infty$, and $\gamma_q^{L_q+1} = +\infty$.

Let $\boldsymbol{\ell} = (\ell_1, \cdots, \ell_Q)^\top$ be a possible value of $\mathbf{z}$. Here, for example, ℓ_1 can be any one of $1, \cdots, L_1$. Clearly,

$$[\mathbf{z} = \boldsymbol{\ell} \mid \mathbf{x} = \mathbf{x}_{(s)}, \mathbf{y}] = \int_{\mathcal{S}} \phi_Q(\mathbf{v} \mid \mathbf{R})\, d\mathbf{v}, \tag{4.5}$$

where $\phi_Q(\cdot \mid \mathbf{R})$ is the Q-dimensional normal density with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$, and

$$\mathcal{S} = \left\{(v_1, \cdots, v_Q) : \nu_{sq}^{\ell_q - 1} < v_q \leq \nu_{sq}^{\ell_q}, q = 1, \cdots, Q\right\},$$

with $\nu_{sq}^{\ell_q} = \gamma_q^{\ell_q} - \tau_{sq} - \boldsymbol{\beta}_q^\top \mathbf{y}$. The joint density $[\mathbf{x}, \mathbf{y}, \mathbf{z}]$ of $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ can thus be written as

$$\begin{aligned}
[\mathbf{x} = \mathbf{x}_{(s)}, \mathbf{y}, \mathbf{z} = \boldsymbol{\ell}] &= [\mathbf{x} = \mathbf{x}_{(s)}; \boldsymbol{\pi}][\mathbf{y} \mid \mathbf{x}_{(s)}; \boldsymbol{\mu}_s, \mathbf{\Sigma}][\mathbf{z} = \boldsymbol{\ell} \mid \mathbf{x}_{(s)}, \mathbf{y}; \boldsymbol{\tau}_s, \mathbf{B}, \mathbf{R}] \\
&= \pi_s \times \phi_C(\mathbf{y} - \boldsymbol{\mu}_s \mid \mathbf{\Sigma}) \int_{\mathcal{S}} \phi_Q(\mathbf{v} \mid \mathbf{R})\, d\mathbf{v}.
\end{aligned} \tag{4.6}$$

It is now possible to formally define the *general mixed-data model* as follows.

Definition 4.1 *The vectors* $\mathbf{x}, \mathbf{y}$ *and* $\mathbf{z}$ *are said to be jointly distributed according to the general mixed-data model if and only if*

(*i*) $\mathbf{x}$ *and* $\mathbf{y}$ *follow the general location model* $GLM(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, *and,*

(*ii*) $\mathbf{y}$ *and* $\mathbf{z}$ *follow the conditional grouped continuous model given* $\mathbf{x} = \mathbf{x}_{(s)}$ $(s = 1, \cdots, S)$ *with parameters given by* $\boldsymbol{\mu}_s$, $\text{vech}(\boldsymbol{\Sigma})$, $\boldsymbol{\gamma}$, $\text{vech}(\mathbf{R})$, $\boldsymbol{\beta} = \text{vec}(\mathbf{B})$, *as previously defined in Definition 3.2, and, in addition,* $\boldsymbol{\tau}^\top = (\boldsymbol{\tau}_1^\top, \cdots, \boldsymbol{\tau}_{S-1}^\top)$.

The parameters of the model are represented by $\boldsymbol{\theta}^\top = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \boldsymbol{\theta}_3^\top)$, *where*

$$\begin{aligned} \boldsymbol{\theta}_1^\top &= (\pi_1, \cdots, \pi_{S-1}), \\ \boldsymbol{\theta}_2^\top &= \left(\boldsymbol{\mu}^\top, \{\text{vech}(\boldsymbol{\Sigma})\}^\top\right), \\ \boldsymbol{\theta}_3^\top &= \left(\boldsymbol{\gamma}^\top, \{\text{vech}(\mathbf{R})\}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\tau}^\top\right). \end{aligned}$$

Further remarks concerning the general mixed-data model are given below.

Remark 4.2.1 *The multivariate probit model in (4.5) is the probit component of the conditional grouped continuous model for state s. It is different from that in the conditional grouped continuous model studied earlier in Chapter 3 because of the presence of the state-specific effects* $\boldsymbol{\tau}_1, \cdots, \boldsymbol{\tau}_{S-1}$. *If* $\boldsymbol{\tau}_1 = \cdots = \boldsymbol{\tau}_{S-1} = \mathbf{0}$ *(i.e.,* $\boldsymbol{\xi}_s^* = \mathbf{D}\mathbf{B}\boldsymbol{\xi}_s$*), then (4.5) reduces to (3.2).*

Remark 4.2.2 *In the case of a single ordinal variable* $(Q = 1)$, *(4.6) reduces to*

$$[\mathbf{x} = \mathbf{x}_{(s)}, \mathbf{y}, Z = \ell] = \pi_s \times \phi_C\left(\mathbf{y} - \boldsymbol{\mu}_s \mid \boldsymbol{\Sigma}\right)\left[\Phi(\nu_s^\ell) - \Phi(\nu_s^{\ell-1})\right], \qquad (4.7)$$

where Φ is the univariate standard normal distribution function. Further, if $\boldsymbol{\sigma}$ is the covariance vector between $\mathbf{y}$ and the latent variable underlying Z, it is easy to see that $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}/\left(1-\boldsymbol{\sigma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}\right)$ and $\tau_s = (\xi_s^* - \boldsymbol{\sigma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}_s)/\left(1-\boldsymbol{\sigma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\sigma}\right)$, $s = 1, \cdots, S-1$. This model includes those studied by Lee and Poon (1986) and Hannan and Tate (1965) as special cases.

Remark 4.2.3 *There are a total of $P = (S-1)+(C+Q)(2S+1)/2+(C+Q)^2/2+L-Q$ independent parameters in the general mixed-data model, $L-Q$ more parameters than in the general location model.*

Remark 4.2.4 *From a block diagonal covariance matrix $\boldsymbol{\Gamma}$ (i.e., $\boldsymbol{\Sigma}_{\mathbf{yy}^*} = \mathbf{0}$), it follows that $\mathbf{B} = \mathbf{0}$ and the probit model in (4.5) reduces to a multi-state grouped continuous model. Anderson and Pemberton's (1985) model is obtained by taking $C = 0$ and $S = 1$.*

Remark 4.2.5 *The general location model is obtained from the general mixed-data model by setting $Q = 0$, and hence, the former may be viewed as a special case of the latter. Similarly, the general mixed-data model reduces to the conditional grouped continuous model when $S = 1$. Therefore, Definition 4.1 unifies these two mixed-data models into a general model.*

Remark 4.2.6 *The state-specific effects $\boldsymbol{\tau}_s$ induce the association between $\mathbf{x}$ and $\mathbf{z}$. The regression effects $\boldsymbol{\beta}_q$ represent the polyserial correlations between $\mathbf{y}$ and $\mathbf{z}$. The standardized cutpoints $\gamma_q^{\ell_q}$ account for the ordinal information contained in $\mathbf{z}$. The associations between the ordinal variables are captured*

by their polychoric correlations $r_{qq'}$. Hence, the general mixed-data model addresses the shortcomings of both the general location and conditional grouped continuous models.

It should be noted that the general mixed-data model was first developed in de Leon and Carrière (2001) to allow for ordinal variables to be incorporated into the general location model.

4.2.1 Case with $C = L = Q = 1$ and $S = 2$: An Example

In this section, the general mixed-data model with two nominal categories ($S = 2$) represented by $\mathbf{x} = (X_1, X_2)^\top$, one ordinal variable Z ($Q = 1$) with two levels so that $L = 1$, and one continuous variable Y ($C = 1$), is explored.

Suppose $\mathbf{x} = \mathbf{x}_{(1)}$ and $\mathbf{x} = \mathbf{x}_{(2)}$ have respective probabilities p and $q = 1 - p$, and the conditional joint distributions of $(Y, Y^*)^\top$ for $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$ are assumed to be bivariate normal with respective mean vectors $(\mu_1, 0)^\top$ and $(\mu_2, 0)^\top$ and common covariance matrix

$$\mathbf{\Gamma} = \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}.$$

Instead of observing Y^*, a dichotomized variable Z is observed such that

$$Z = \begin{cases} 1 & \text{if } Y^* \leq \alpha \\ 2 & \text{otherwise} \end{cases},$$

where α is an unknown cutpoint. In this case, the model parameter vector becomes $\boldsymbol{\theta}^\top = (p, \mu_1, \mu_2, \sigma^2, \gamma, \beta, \tau)$, where $\gamma = \alpha/\sqrt{1-\rho^2} - \beta\mu_2$, $\beta = \rho/(\sigma\sqrt{1-\rho^2})$, and $\tau = -\beta\xi$, where $\xi = \mu_1 - \mu_2$.

Direct calculations (see also Tallis, 1961) show that $\mathrm{E}(Y^* \mid \mathbf{x} = \mathbf{x}_{(1)}, Z = 2) = \mathrm{E}(Y^* \mid \mathbf{x} = \mathbf{x}_{(2)}, Z = 2) = \phi(\alpha)/\Phi(-\alpha)$ and $\mathrm{E}(Y^* \mid \mathbf{x} = \mathbf{x}_{(1)}, Z = 1) =$

$\mathrm{E}(Y^* \mid \mathbf{x} = \mathbf{x}_{(2)}, Z = 1) = -\phi(\alpha)/\Phi(\alpha)$. Similarly, it can be shown that the following hold as well:

$$\begin{aligned}
\mathrm{var}(Y^* \mid \mathbf{x} = \mathbf{x}_{(1)}, Z = 1) &= 1 - \frac{\alpha\phi(\alpha)}{\Phi(\alpha)} - \left[\frac{\phi(\alpha)}{\Phi(\alpha)}\right]^2 \\
&= \mathrm{var}(Y^* \mid \mathbf{x} = \mathbf{x}_{(2)}, Z = 1), \\
\mathrm{var}(Y^* \mid \mathbf{x} = \mathbf{x}_{(1)}, Z = 2) &= 1 + \frac{\alpha\phi(\alpha)}{\Phi(-\alpha)} - \left[\frac{\phi(\alpha)}{\Phi(-\alpha)}\right]^2 \\
&= \mathrm{var}(Y^* \mid \mathbf{x} = \mathbf{x}_{(2)}, Z = 2).
\end{aligned}$$

Noting that, given $\mathbf{x} = \mathbf{x}_{(s)}$, $Y = \mu_s + \rho\sigma Y^* + \varepsilon$, with $\mathrm{E}(\varepsilon) = 0$, $\mathrm{var}(\varepsilon) = \sigma^2(1-\rho^2)$ and $\mathrm{cov}(Y^*, \varepsilon) = 0$, it is easy to see that

$$\begin{aligned}
\mathrm{E}(Y \mid \mathbf{x} = \mathbf{x}_{(1)}, Z = 1) &= \mu_1 + \frac{\rho\sigma\phi(\alpha)}{\Phi(-\alpha)} \\
&= \mathrm{E}(Y \mid \mathbf{x} = \mathbf{x}_{(1)}, Z = 2), \\
\mathrm{E}(Y \mid \mathbf{x} = \mathbf{x}_{(2)}, Z = 1) &= \mu_2 - \frac{\rho\sigma\phi(\alpha)}{\Phi(\alpha)} \\
&= \mathrm{E}(Y \mid \mathbf{x} = \mathbf{x}_{(2)}, Z = 2),
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{var}(Y \mid \mathbf{x} = \mathbf{x}_{(1)}, Z = 1) &= \sigma^2(1-\rho^2) + \rho^2\sigma^2\left\{1 - \frac{\alpha\phi(\alpha)}{\Phi(\alpha)} - \left[\frac{\phi(\alpha)}{\Phi(\alpha)}\right]^2\right\} \\
&= \mathrm{var}(Y \mid \mathbf{x} = \mathbf{x}_{(2)}, Z = 1), \\
\mathrm{var}(Y \mid \mathbf{x} = \mathbf{x}_{(1)}, Z = 2) &= \sigma^2(1-\rho^2) + \rho^2\sigma^2\left\{1 + \frac{\alpha\phi(\alpha)}{\Phi(-\alpha)} - \left[\frac{\phi(\alpha)}{\Phi(-\alpha)}\right]^2\right\} \\
&= \mathrm{var}(Y \mid \mathbf{x} = \mathbf{x}_{(2)}, Z = 2).
\end{aligned}$$

Suppose now that the binary variable Z is taken as nominal and a general location model is assumed for the distribution of $(\mathbf{x}^\top, Y, Z)^\top$. In this case, the following remarks apply.

Remark 4.2.7 *The vector* $(\mathbf{x}^\top, Z)^\top$ *defines a* 2×2 *contingency table for which* Y *is assumed to have a normal distribution whose mean varies across states but with constant variance. This is certainly not the case with the general mixed-data model described above, where the means of* Y *for states* $(\mathbf{x} = \mathbf{x}_{(1)}, Z = 1)$ *and* $(\mathbf{x} = \mathbf{x}_{(1)}, Z = 2)$ *are the same, and so are those for states* $(\mathbf{x} = \mathbf{x}_{(2)}, Z = 1)$ *and* $(\mathbf{x} = \mathbf{x}_{(2)}, Z = 2)$.

Remark 4.2.8 *While the general location model assumes homogeneity of the state variances, the general mixed-data model does not. As exhibited above, the variance for states* $(\mathbf{x} = \mathbf{x}_{(1)}, Z = 1)$ *and* $(\mathbf{x} = \mathbf{x}_{(2)}, Z = 1)$ *is different from that for* $(\mathbf{x} = \mathbf{x}_{(1)}, Z = 2)$ *and* $(\mathbf{x} = \mathbf{x}_{(2)}, Z = 2)$.

Remark 4.2.9 *It should be noted that the general mixed-data model reduces to the general location model with two states for* $\alpha = 0$, *as states* $(\mathbf{x} = \mathbf{x}_{(s)}, Z = 1)$ *and* $(\mathbf{x} = \mathbf{x}_{(s)}, Z = 2)$, $s = 1, 2$, *are collapsed. This is so because it was assumed in the model above that* $\mathrm{E}(Y^*) = 0$ *and* $\mathrm{var}(Y^*) = 1$, *which do not depend on* $\mathbf{x}$.

The next section discusses maximum likelihood estimation for the general mixed-data model developed above.

4.3 Maximum Likelihood Estimation

Suppose a mixed-variable random sample $(\mathbf{x}_i^\top, \mathbf{y}_i^\top, \mathbf{z}_i^\top)^\top$, $i = 1, \cdots, N$, is observed. For convenience of development, the cases (i) $Q = 1$ and (ii) $Q \geq 2$ are considered separately. An example of the data layout is given in Figure 4.2 for the case $Q = S = 2, L_1 = L_2 = 1$ and $C \geq 1$.

4.3.1 Case of a Single Ordinal Variable ($Q = 1$)

Define the sets $\mathcal{A}(s) = \{i \mid \mathbf{x}_i = \mathbf{x}_{(s)}\}$ and $\mathcal{B}(\ell) = \{i \mid Z_i = \ell\}$, $s = 1, \cdots, S$, $\ell = 1, \cdots, L+1$. Using (4.7) and § 3.3, it follows that

$$\begin{aligned} \mathcal{L} &= \left[(1 - \pi_1 - \cdots - \pi_{S-1})^{n_S} \prod_{s=1}^{S-1} \pi_s^{n_s}\right] \phi_C^N(\mathbf{y}_1, \cdots, \mathbf{y}_N \mid \boldsymbol{\theta}_2) \\ &\quad \times \prod_{s=1}^{S} \prod_{\ell=1}^{L+1} \prod_{i(s,\ell)} \left[\Phi(\nu_{i(s,\ell)}^{\ell}) - \Phi(\nu_{i(s,\ell)}^{\ell-1})\right], \end{aligned} \tag{4.8}$$

where

$$\begin{aligned} \phi_C^N(\mathbf{y}_1, \cdots, \mathbf{y}_N \mid \boldsymbol{\theta}_2) &= (2\pi)^{-N/2} \mid \boldsymbol{\Sigma} \mid^{-N/2} \\ &\quad \times \exp\left[-\frac{1}{2} \sum_{s=1}^{S} \sum_{i(s)} (\mathbf{y}_{i(s)} - \boldsymbol{\mu}_s)^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{y}_{i(s)} - \boldsymbol{\mu}_s)\right], \end{aligned}$$

$n_s = \sum_{i=1}^{N} x_{is}$, and the indices "$i(s)$" and "$i(s,\ell)$" come from $\mathcal{A}(s)$ and $\mathcal{A}(s) \cap \mathcal{B}(\ell)$, respectively, $s = 1, \cdots, S$; $\ell = 1, \cdots, L+1$. Note that $\mathbf{y}_{i(s)} \in \{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$ and $\mathbf{y}_{i(s,\ell)} \in \{\mathbf{y}_{1(s)}, \cdots, \mathbf{y}_{n_s(s)}\}$. Here, $\gamma^0 = -\infty$ and $\gamma^{L+1} = +\infty$. The log-likelihood is given by

$$\log \mathcal{L} = \ell_1(\boldsymbol{\theta}_1) + \ell_2(\boldsymbol{\theta}_2) + \ell_3(\boldsymbol{\theta}_3), \tag{4.9}$$

where

$$\begin{aligned} \ell_1(\boldsymbol{\theta}_1) &= \sum_{s=1}^{S-1} n_s \log \pi_s, \\ \ell_2(\boldsymbol{\theta}_2) &= -\frac{N}{2} \log 2\pi - \frac{N}{2} \log \mid \boldsymbol{\Sigma} \mid - \frac{1}{2} \sum_{s=1}^{S} \sum_{i(s)} (\mathbf{y}_{i(s)} - \boldsymbol{\mu}_s)^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{y}_{i(s)} - \boldsymbol{\mu}_s), \\ \ell_3(\boldsymbol{\theta}_3) &= \sum_{s=1}^{S} \sum_{\ell=1}^{L+1} \sum_{i(s,\ell)} \log\left[\Phi(\nu_{i(s,\ell)}^{\ell}) - \Phi(\nu_{i(s,\ell)}^{\ell-1})\right]. \end{aligned}$$

Noting that the space for $\boldsymbol{\theta}$ is simply the product of the individual spaces for $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$, the MLE $\widehat{\boldsymbol{\theta}}^{\top} = (\widehat{\boldsymbol{\theta}}_1^{\top}, \widehat{\boldsymbol{\theta}}_2^{\top}, \widehat{\boldsymbol{\theta}}_3^{\top})$ is found by maximizing $\ell_1(\boldsymbol{\theta}_1)$,

$\ell_2(\boldsymbol{\theta}_2)$ and $\ell_3(\boldsymbol{\theta}_3)$ separately. The result for $\boldsymbol{\theta}_1$ is the usual MLE for a multinomial model given by $\widehat{\pi}_s = n_s/N$, while that for $\boldsymbol{\theta}_2$ is given by $\widehat{\boldsymbol{\mu}}_s = \overline{\mathbf{y}}_s = \sum_{i(s)} \mathbf{y}_{i(s)}/n_s$, and the unique elements of $\widehat{\boldsymbol{\Sigma}} = \mathbf{S} = \sum_{s=1}^{S} \sum_{i(s)} (\mathbf{y}_{i(s)} - \overline{\mathbf{y}}_s)(\mathbf{y}_{i(s)} - \overline{\mathbf{y}}_s)^\top / N$, where $\overline{\mathbf{y}}_s$ is the sth state mean, $s = 1, \cdots, S$. See § 2.2.2.

The MLE $\widehat{\boldsymbol{\theta}}_3$ of $\boldsymbol{\theta}_3$ is obtained using an iterative technique such as the Newton-Raphson method. Let $\mathbf{s}(\boldsymbol{\theta}_3) = \partial \ell_3(\boldsymbol{\theta}_3)/\partial \boldsymbol{\theta}_3$ and $\mathbf{H}(\boldsymbol{\theta}_3) = \partial^2 \ell_3(\boldsymbol{\theta}_3)/\partial \boldsymbol{\theta}_3 \partial \boldsymbol{\theta}_3^\top$ be the score vector and the Hessian matrix of $\boldsymbol{\theta}_3$, respectively.

Lemma 4.1 *Consider the general mixed-data model with $Q = 1$. Then, the elements of the score vector $\mathbf{s}(\boldsymbol{\theta}_3)$ are given by the following:*

$$\begin{aligned}
\frac{\partial \ell_3(\boldsymbol{\theta}_3)}{\partial \boldsymbol{\beta}} &= \sum_{s=1}^{S} \sum_{\ell=1}^{L+1} \sum_{i(s,\ell)} \frac{-\mathbf{y}_{i(s,\ell)} \delta^{\ell}_{i(s,\ell)}}{\Delta^{\ell}_{i(s,\ell)}}, \\
\frac{\partial \ell_3(\boldsymbol{\theta}_3)}{\partial \gamma^{\ell}} &= \sum_{s=1}^{S} \sum_{i(s,\ell)} \frac{\phi^{\ell}_{i(s,\ell)}}{\Delta^{\ell}_{i(s,\ell)}} - \sum_{s=1}^{S} \sum_{i(s,\ell+1)} \frac{\phi^{\ell}_{i(s,\ell+1)}}{\Delta^{\ell+1}_{i(s,\ell+1)}}, \quad (\ell = 1, \cdots, L) \\
\frac{\partial \ell_3(\boldsymbol{\theta}_3)}{\partial \tau_s} &= \sum_{\ell=1}^{L+1} \sum_{i(s,\ell)} \frac{-\delta^{\ell}_{i(s,\ell)}}{\Delta^{\ell}_{i(s,\ell)}}, \quad (s = 1, \cdots, S-1)
\end{aligned}$$

where $\Delta^{\ell}_{i(s,\ell)} \equiv \Phi^{\ell}_{i(s,\ell)} - \Phi^{\ell-1}_{i(s,\ell)}$ and $\delta^{\ell}_{i(s,\ell)} \equiv \phi^{\ell}_{i(s,\ell)} - \phi^{\ell-1}_{i(s,\ell)}$, with $\Phi^{\ell}_{i(s,\ell)} \equiv \Phi(\nu^{\ell}_{i(s,\ell)})$ and $\phi^{\ell}_{i(s,\ell)} \equiv \phi(\nu^{\ell}_{i(s,\ell)})$.

In addition, the elements of the Hessian matrix $\mathbf{H}(\boldsymbol{\theta}_3)$ are given by the following:

$$\frac{\partial^2 \ell_3(\boldsymbol{\theta}_3)}{\partial \gamma^{\ell'} \partial \gamma^{\ell}} = \begin{cases}
\sum_{s=1}^{S} \sum_{i(s,\ell)} \frac{\phi^{\ell-1}_{i(s,\ell)} \phi^{\ell}_{i(s,\ell)}}{(\Delta^{\ell}_{i(s,\ell)})^2} & \ell' = \ell - 1, \\
\sum_{s=1}^{S} \sum_{i(s,\ell)} \frac{\phi^{\ell}_{i(s,\ell)}}{(\Delta^{\ell}_{i(s,\ell)})^2} (\nu^{\ell}_{i(s,\ell)} \Delta^{\ell}_{i(s,\ell)} - \phi^{\ell}_{i(s,\ell)}) & \\
\quad - \sum_{s=1}^{S} \sum_{i(s,\ell+1)} \frac{\phi^{\ell}_{i(s,\ell+1)}}{(\Delta^{\ell+1}_{i(s,\ell+1)})^2} (\nu^{\ell}_{i(s,\ell+1)} \Delta^{\ell+1}_{i(s,\ell+1)} - \phi^{\ell}_{i(s,\ell+1)}) & \ell' = \ell, \\
\sum_{s=1}^{S} \sum_{i(s,\ell+1)} \frac{\phi^{\ell}_{i(s,\ell+1)} \phi^{\ell+1}_{i(s,\ell+1)}}{(\Delta^{\ell+1}_{i(s,\ell+1)})^2} & \ell' = \ell + 1, \\
0 & \textit{otherwise},
\end{cases}$$

$$
\begin{aligned}
\frac{\partial^2 \ell_3(\boldsymbol{\theta}_3)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} &= \sum_{s=1}^{S}\sum_{\ell=1}^{L+1}\sum_{i(s,\ell)} \frac{\mathbf{y}_{i(s,\ell)}\mathbf{y}_{i(s,\ell)}^\top}{(\Delta_{i(s,\ell)}^{\ell})^2}(\delta_{i(s,\ell)}^{\ell} + \delta_{\nu,i(s,\ell)}^{\ell}\Delta_{i(s,\ell)}^{\ell}), \\
\frac{\partial^2 \ell_3(\boldsymbol{\theta}_3)}{\partial^2\tau_s} &= \sum_{\ell=1}^{L+1}\sum_{i(s,\ell)} \left[\frac{\delta_{\nu,i(s,\ell)}^{\ell}}{\Delta_{i(s,\ell)}^{\ell}} - \left(\frac{\delta_{i(s,\ell)}^{\ell}}{\Delta_{i(s,\ell)}^{\ell}}\right)^2\right], \\
\frac{\partial^2 \ell_3(\boldsymbol{\theta}_3)}{\partial\boldsymbol{\beta}^\top\partial\gamma^{\ell}} &= \sum_{s=1}^{S}\sum_{i(s,\ell)} \frac{\phi_{i(s,\ell)}^{\ell}\mathbf{y}_{i(s,\ell)}^\top}{(\Delta_{i(s,\ell)}^{\ell})^2}\left(\delta_{i(s,\ell)}^{\ell} - \nu_{i(s,\ell)}^{\ell}\Delta_{i(s,\ell)}^{\ell}\right) \\
&\quad - \sum_{s=1}^{S}\sum_{i(s,\ell+1)} \frac{\phi_{i(s,\ell+1)}^{\ell}\mathbf{y}_{i(s,\ell+1)}^\top}{(\Delta_{i(s,\ell+1)}^{\ell+1})^2}(\delta_{i(s,\ell+1)}^{\ell} - \nu_{i(s,\ell+1)}^{\ell}\Delta_{i(s,\ell+1)}^{\ell+1}), \\
\frac{\partial^2 \ell_3(\boldsymbol{\theta}_3)}{\partial\boldsymbol{\beta}^\top\partial\tau_s} &= \sum_{\ell=1}^{L+1}\sum_{i(s,\ell)} \frac{\mathbf{y}_{i(s,\ell)}^\top}{(\Delta_{i(s,\ell)}^{\ell})^2}\left(\delta_{i(s,\ell)}^{\ell} + \delta_{\nu,i(s,\ell)}^{\ell}\Delta_{i(s,\ell)}^{\ell}\right), \\
\frac{\partial^2 \ell_3(\boldsymbol{\theta}_3)}{\partial\tau_s\partial\gamma^{\ell}} &= \sum_{s=1}^{S}\sum_{i(s,\ell)} \frac{\phi_{i(s,\ell)}^{\ell}}{(\Delta_{i(s,\ell)}^{\ell})^2}\left(\delta_{i(s,\ell)}^{\ell} - \nu_{i(s,\ell)}^{\ell}\Delta_{i(s,\ell)}^{\ell}\right) \\
&\quad - \sum_{s=1}^{S}\sum_{i(s,\ell+1)} \frac{\phi_{i(s,\ell+1)}^{\ell}}{(\Delta_{i(s,\ell+1)}^{\ell+1})^2}\left(\delta_{i(s,\ell+1)}^{\ell+1} - \nu_{i(s,\ell+1)}^{\ell}\Delta_{i(s,\ell+1)}^{\ell+1}\right),
\end{aligned}
$$

where $\delta_{\nu,i(s,\ell)}^{\ell} \equiv \nu_{i(s,\ell)}^{\ell}\phi_{i(s,\ell)}^{\ell} - \nu_{i(s,\ell)}^{\ell-1}\phi_{i(s,\ell)}^{\ell-1}$. *Note also that* $\partial^2\ell_3(\boldsymbol{\theta}_3)/\partial\tau_s\partial\tau_{s'} = 0$, $\partial^2\ell_3(\boldsymbol{\theta}_3)/\partial\gamma^0\partial\gamma^1 = 0$, *and* $\partial^2\ell_3(\boldsymbol{\theta}_3)/\partial\gamma^{L+1}\partial\gamma^L = 0$.

Proof. The proof is straightforward and hence, is omitted. See also Lee and Poon (1986).

□

From Lemma 4.1, $\boldsymbol{\theta}_3$ can be estimated by the Newton-Raphson method via the updating formula

$$
\widehat{\boldsymbol{\theta}}_3^{(t+1)} = \widehat{\boldsymbol{\theta}}_3^{(t)} - \mathbf{H}^{-1}(\boldsymbol{\theta}_3)\mathbf{s}(\boldsymbol{\theta}_3)\Big|_{\boldsymbol{\theta}_3=\widehat{\boldsymbol{\theta}}_3^{(t)}}, \quad t = 0, 1, \cdots, \tag{4.10}
$$

where it is assumed that $\mathbf{H}(\widehat{\boldsymbol{\theta}}_3^{(t)}) > \mathbf{0}$, with $\widehat{\boldsymbol{\theta}}_3^{(t)}$ the estimate at the tth iteration. Iterations continue until convergence is attained, i.e., $\mid \ell_3(\widehat{\boldsymbol{\theta}}_3^{(t+1)}) - \ell_3(\widehat{\boldsymbol{\theta}}_3^{(t)}) \mid \leq \varepsilon$,

where $\varepsilon > 0$ is a pre-specified tolerance. For the initial estimate $\widehat{\boldsymbol{\theta}}_3^{(0)}$, the following are suggested:

$$\begin{aligned}
\widehat{\gamma}^{\ell,(0)} &= \Phi^{-1}\left(\sum_{s=1}^{S}\sum_{\ell'=1}^{\ell}\frac{n_{s\ell'}}{N}\right), \\
\widehat{\boldsymbol{\beta}}^{(0)} &= \frac{\mathbf{S}^{-1}\widetilde{\boldsymbol{\sigma}}}{\sqrt{1-\widetilde{\boldsymbol{\sigma}}^{\top}\mathbf{S}^{-1}\widetilde{\boldsymbol{\sigma}}}}, \\
\widehat{\tau}_s^{(0)} &= \frac{(\overline{Y}_s^{*}-\overline{Y}_S^{*})-\widetilde{\boldsymbol{\sigma}}^{\top}\mathbf{S}^{-1}(\overline{\mathbf{y}}_s-\overline{\mathbf{y}}_S)}{\sqrt{1-\widetilde{\boldsymbol{\sigma}}^{\top}\mathbf{S}^{-1}\widetilde{\boldsymbol{\sigma}}}},
\end{aligned} \tag{4.11}$$

where $\overline{Y}_s^{*}$ is an estimate of the sth latent mean based on frequency data, $\overline{Y}_S^{*}$ and $\overline{\mathbf{y}}_S$ are the reference state means, $n_{s\ell}$ is the number of observations in state s such that $Z_i = \ell$, and the cth element of $\widetilde{\mathrm{cov}}(\mathbf{y}, Z) = \widetilde{\boldsymbol{\sigma}}$ is given by $\widetilde{\sigma}_c = \widetilde{r}_c\sqrt{s_{cc}}$, the sample covariance between the cth element of $\mathbf{y}$ and the latent variable underlying Z. Here, $\widetilde{r}_c$ is the sample *point polyserial correlation* (Drasgow, 1986; Lee and Poon, 1986) and s_{cc} is the cth diagonal element of $\mathbf{S}$.

4.3.2 General Case ($Q \geq 2$)

The case when $Q \geq 2$ is now considered. Define $\mathcal{B}(\ell_1, \cdots, \ell_Q) = \{i \mid Z_{iq} = \ell_q, \ell_q = 1, \cdots, L_q + 1; q = 1, \cdots, Q\}$. Following the approach used for the grouped continuous model in Chapter 3 (see also Poon and Lee, 1987), it can be shown that $\ell_3(\boldsymbol{\theta}_3)$ is given by

$$\begin{aligned}
\ell_3(\boldsymbol{\theta}_3) = \sum_{s=1}^{S}\sum_{\ell_1=1}^{L_1+1}\cdots\sum_{\ell_Q=1}^{L_Q+1}\sum_{i(s,\boldsymbol{\ell})}\log\Bigg[\sum_{\epsilon_1=0}^{1}\cdots \\
\sum_{\epsilon_Q=0}^{1}(-1)^{\sum_{q=1}^{Q}\epsilon_q+Q}\Phi_Q\left(\cdots, \nu_{i(s,\boldsymbol{\ell})q}^{\ell_q-\epsilon_q}, \cdots \mid \mathbf{R}\right)\Bigg],
\end{aligned}$$

where $\Phi_Q(\cdot \mid \mathbf{R})$ is the Q-dimensional normal distribution function with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$, and $\nu_{i(s,\boldsymbol{\ell})q}^{\ell_q} = \gamma_q^{\ell_q} - \tau_{sq} - \boldsymbol{\beta}_q^{\top}\mathbf{y}_{i(s,\boldsymbol{\ell})}$. The index

"$i(s,\boldsymbol{\ell})$" comes from $\mathcal{A}(s) \cap \mathcal{B}(\ell_1, \cdots, \ell_Q)$, and refers to the ith unit in state s such that $Z_1 = \ell_1, \cdots, Z_Q = \ell_Q$. Note that $\nu^0_{i(s,\boldsymbol{\ell})q} = -\infty$ and $\nu^{L_q+1}_{i(s,\boldsymbol{\ell})q} = +\infty$, $q = 1, \cdots, Q$.

As in the case $Q = 1$, the MLE $\widehat{\boldsymbol{\theta}}_3$ can be obtained via the updating formula (4.10), using a different score vector and Hessian matrix, or by the *Fletcher-Powell algorithm* (Fletcher and Powell, 1964) suggested by Drasgow (1986) and advocated by Poon and Lee (1987). The advantage of the Fletcher-Powell algorithm is that it only requires the score vector and an initial positive definite matrix, usually taken as the identity matrix.

In what follows, define $\Delta^{\ell_q}_{\phi,i(s,\boldsymbol{\ell}),Q-1} \equiv \phi^{\ell_q}_{i(s,\boldsymbol{\ell})q} \Phi^{(q)}_{i(s,\boldsymbol{\ell}),Q-1} - \phi^{\ell_q-1}_{i(s,\boldsymbol{\ell})q} \Phi^{(q-1)}_{i(s,\boldsymbol{\ell}),Q-1}$ and $\phi^{\ell_q,\ell_{q'}}_{i(s,\boldsymbol{\ell})qq'} \equiv \phi_2(\nu^{\ell_q}_{i(s,\boldsymbol{\ell})q}, \nu^{\ell_{q'}}_{i(s,\boldsymbol{\ell})q'} \mid r_{qq'})$, where $\phi^{\ell_q}_{i(s,\boldsymbol{\ell})q} \equiv \phi(\nu^{\ell_q}_{i(s,\boldsymbol{\ell})q})$ and $\phi_2(\cdot,\cdot \mid r_{qq'})$ is the standard bivariate normal density with correlation coefficient $r_{qq'}$, and

$$\begin{aligned}
\Phi^{(q)}_{i(s,\boldsymbol{\ell}),Q-1} &\equiv \Phi_{Q-1}\left(\cdots, \frac{\nu^{\ell_{q'}-\epsilon_{q'}}_{i(s,\boldsymbol{\ell})q'} - r_{qq'}\nu^{\ell_q-\epsilon_q}_{i(s,\boldsymbol{\ell})q}}{\sqrt{1-r^2_{qq'}}}, \cdots; q' \neq q \mid \mathbf{R}_{\cdot q}\right), \\
\Phi^{(q,q')}_{i(s,\boldsymbol{\ell}),Q-2} &\equiv \Phi_{Q-2}\left(\cdots, \nu^{\ell_{q''}-\epsilon_{q''}}_{i(s,\boldsymbol{\ell})q''} - h_{q''}, \cdots; q'' \neq q; q'' \neq q' \mid \mathbf{R}_{\cdot qq'}\right),
\end{aligned}$$

are the $(Q-1)$- and $(Q-2)$-dimensional normal distribution functions with $\mathbf{0}$ means and covariance matrices equal, respectively, to the partial correlation matrix $\mathbf{R}_{\cdot q}$, given Y^*_q, and the partial correlation matrix $\mathbf{R}_{\cdot qq'}$, given Y^*_q and $Y^*_{q'}$, with

$$h_{q''} = \frac{r_{qq''} - r_{q'q''}r_{qq'}}{\sqrt{1-r^2_{qq'}}}\nu^{\ell_q}_{i(s,\boldsymbol{\ell})q} + \frac{r_{q'q''} - r_{qq''}r_{qq'}}{\sqrt{1-r^2_{qq'}}}\nu^{\ell_{q'}}_{i(s,\boldsymbol{\ell})q'}.$$

Finally, let

$$A_{i(s,\boldsymbol{\ell})} = \sum_{\epsilon_1=0}^{1} \cdots \sum_{\epsilon_Q=0}^{1} (-1)^{\sum_{q=1}^{Q} \epsilon_q + Q} \Phi_Q \left(\cdots, \nu_{i(s,\boldsymbol{\ell})q}^{\ell_q - \epsilon_q}, \cdots \mid \mathbf{R} \right).$$

Lemma 4.2 *Consider the general mixed-data model with $Q \geq 2$. Then, the elements of the score vector $\mathbf{s}(\boldsymbol{\theta}_3)$ are given by the following:*

$$\begin{aligned}
\frac{\partial \ell_3(\boldsymbol{\theta}_3)}{\partial \gamma_q^{\ell_q}} &= \sum_{s=1}^{S} \sum_{\ell_1=1}^{L_1+1} \cdots \sum_{\ell_{q-1}=1}^{L_{q-1}+1} \sum_{\ell_{q+1}=1}^{L_{q+1}+1} \cdots \sum_{\ell_Q=1}^{L_Q+1} \sum_{i(s,\boldsymbol{\ell})} A_{i(s,\boldsymbol{\ell})}^{-1} \\
&\quad \sum_{\epsilon_1=0}^{1} \cdots \sum_{\epsilon_{q-1}=0}^{1} \sum_{\epsilon_{q+1}=0}^{1} \cdots \sum_{\epsilon_Q=0}^{1} (-1)^{\sum_{q' \neq q} \epsilon_{q'} + Q} \phi_{i(s,\boldsymbol{\ell})q}^{\ell_q} \Phi_{i(s,\boldsymbol{\ell}),Q-1}^{(q)}, \\
\frac{\partial \ell_3(\boldsymbol{\theta}_3)}{\partial r_{qq'}} &= \sum_{s=1}^{S} \sum_{\ell_1=1}^{L_1+1} \cdots \sum_{\ell_Q=1}^{L_Q+1} \sum_{i(s,\boldsymbol{\ell})} A_{i(s,\boldsymbol{\ell})}^{-1} \sum_{\epsilon_1=0}^{1} \cdots \sum_{\epsilon_Q=0}^{1} \\
&\quad (-1)^{\sum_{q''=1}^{Q} \epsilon_{q''} + Q} \phi_{i(s,\boldsymbol{\ell})qq'}^{\ell_q, \ell_{q'}} \Phi_{i(s,\boldsymbol{\ell}),Q-2}^{(q,q')}, \\
\frac{\partial \ell_3(\boldsymbol{\theta}_3)}{\partial \boldsymbol{\beta}_q} &= \sum_{s=1}^{S} \sum_{\ell_1=1}^{L_1+1} \cdots \sum_{\ell_{q-1}=1}^{L_{q-1}+1} \sum_{\ell_{q+1}=1}^{L_{q+1}+1} \cdots \sum_{\ell_Q=1}^{L_Q+1} \sum_{i(s,\boldsymbol{\ell})} A_{i(s,\boldsymbol{\ell})}^{-1} \\
&\quad \cdots \sum_{\epsilon_1=0}^{1} \sum_{\epsilon_{q-1}=0}^{1} \sum_{\epsilon_{q+1}=0}^{1} \cdots \sum_{\epsilon_Q=0}^{1} (-1)^{\sum_{q' \neq q} \epsilon_{q'} + Q + 1} \mathbf{y}_{i(s,\boldsymbol{\ell})} \Delta_{\phi,i(s,\boldsymbol{\ell}),Q-1}^{\ell_q}, \\
\frac{\partial \ell_3(\boldsymbol{\theta}_3)}{\partial \tau_{sq}} &= \sum_{\ell_1=1}^{L_1+1} \cdots \sum_{\ell_{q-1}=1}^{L_{q-1}+1} \sum_{\ell_{q+1}=1}^{L_{q+1}+1} \cdots \sum_{\ell_Q=1}^{L_Q+1} \sum_{i(s,\boldsymbol{\ell})} A_{i(s,\boldsymbol{\ell})}^{-1} \\
&\quad \sum_{\epsilon_1=0}^{1} \cdots \sum_{\epsilon_{q-1}=0}^{1} \sum_{\epsilon_{q+1}=0}^{1} \cdots \sum_{\epsilon_Q=0}^{1} (-1)^{\sum_{q' \neq q} \epsilon_{q'} + Q + 1} \Delta_{\phi,i(s,\boldsymbol{\ell}),Q-1}^{\ell_q}.
\end{aligned}$$

Proof. From properties of the multivariate normal distribution (Anderson, 1984, p. 35), it can be shown that

$$\begin{aligned}
\Phi_Q \left(\cdots, \nu_{i(s,\boldsymbol{\ell})q}^{\ell_q - \epsilon_q}, \cdots \mid \mathbf{R} \right) &= \int_{-\infty}^{\nu_{i(s,\boldsymbol{\ell})q}^{\ell_q - \epsilon_q}} \phi(v_q) \\
&\quad \times \Phi_{i(s,\boldsymbol{\ell}),Q-1}^{(q)} \left(\cdots, \frac{\nu_{i(s,\boldsymbol{\ell})q'}^{\ell_{q'} - \epsilon_{q'}} - r_{qq'} v_q}{\sqrt{1 - r_{qq'}^2}}, \cdots \mid v_q, \mathbf{R}_{\cdot q} \right) \mathrm{d}v_q,
\end{aligned}$$

for $q = 1, \cdots, Q$, where $q \neq q'$, so that

$$\frac{\partial}{\partial \nu_{i(s,\boldsymbol{\ell})q}^{\ell_q - \epsilon_q}} \Phi_Q \left(\cdots, \nu_{i(s,\boldsymbol{\ell})q}^{\ell_q - \epsilon_q}, \cdots \mid \mathbf{R} \right) = \phi(\nu_{i(s,\boldsymbol{\ell})q}^{\ell_q - \epsilon_q}) \Phi_{i(s,\boldsymbol{\ell}),Q-1}^{(q)}.$$

The rest of the proof is now straightforward and follows from standard results on vector differentiation (see, e.g., McDonald and Swaminathan, 1973) and the reduction formula of Plackett (1954). See also Poon and Lee (1987).

□

The following initial values are suggested for the Fletcher-Powell algorithm:

$$\begin{aligned} \widehat{\gamma}_q^{\ell_q,(0)} &= \Phi^{-1}\left(\sum_{s=1}^{S} \sum_{\ell_q'=1}^{\ell_q} \frac{n_{s\ell_q'}}{N} \right), \\ \widehat{\boldsymbol{\beta}}_q^{(0)} &= \frac{\mathbf{S}^{-1}\widetilde{\boldsymbol{\sigma}}_q}{\sqrt{1 - \widetilde{\boldsymbol{\sigma}}_q^{\top}\mathbf{S}^{-1}\widetilde{\boldsymbol{\sigma}}_q}}, \\ \widehat{\tau}_{sq}^{(0)} &= \frac{(\overline{Y}_{sq}^{*} - \overline{Y}_{Sq}^{*}) - \widetilde{\boldsymbol{\sigma}}_q^{\top}\mathbf{S}^{-1}(\overline{\mathbf{y}}_s - \overline{\mathbf{y}}_S)}{\sqrt{1 - \widetilde{\boldsymbol{\sigma}}_q^{\top}\mathbf{S}^{-1}\widetilde{\boldsymbol{\sigma}}_q}}, \end{aligned} \tag{4.12}$$

where $\overline{Y}_{sq}^{*}$ and $\overline{Y}_{Sq}$ are estimates of the latent means of the sth and Sth states, $n_{s\ell_q'}$ is the number of observations in state s such that $Z_q = \ell_q'$, $\widetilde{\boldsymbol{\sigma}}_q$ is the sample covariance vector between $\mathbf{y}$ and Z_q, the cth element of which is given by $\widetilde{\sigma}_{qc} = \tilde{r}_{qc}\sqrt{s_{cc}}$, where $\tilde{r}_{qc}$ is the sample point polyserial correlation between the cth element of $\mathbf{y}$ and Z_q.

4.4 Maximum Pairwise Likelihood Estimation

An alternative to maximum likelihood estimation of $\boldsymbol{\theta}$ is presented in this section. The *pairwise likelihood* approach (Kuk and Nott, 2000; Nott and

Rydén, 1999), adopted in Chapter 3, entails specifying a *pseudo-likelihood* based on

$$[\mathbf{x};\boldsymbol{\theta}_1][\mathbf{y} \mid \mathbf{x};\boldsymbol{\theta}_2] \prod_{q<q'} [Z_q, Z_{q'} \mid \mathbf{x}, \mathbf{y};\boldsymbol{\theta}_3]. \tag{4.13}$$

Given a sample, the *pairwise likelihood function* is taken as the product of (4.13) (or sum, in the case of pairwise log-likelihoods) over the sample. Lindsay (1988) refers to this pseudo-likelihood function as a *composite likelihood* and lays down its theory in a very general framework (for an application, see, e.g., Heagerty and Lele, 1998). In general, the composite likelihood approach provides a pseudo-likelihood function from which an estimating equation can be obtained. In the present context, the estimating equation is obtained by maximizing the pairwise log-likelihood function

$$\ell^p = \ell_1^p(\boldsymbol{\theta}_1) + \ell_2^p(\boldsymbol{\theta}_2) + \ell_3^p(\boldsymbol{\theta}_3), \tag{4.14}$$

where $\ell_1^p(\boldsymbol{\theta}_1) = \ell_1(\boldsymbol{\theta}_1)$, $\ell_2^p(\boldsymbol{\theta}_2) = \ell_2(\boldsymbol{\theta}_2)$, and

$$\begin{aligned}\ell_3^p(\boldsymbol{\theta}_3) \;=\; & \sum_{s=1}^{S} \sum_{q<q'} \sum_{\ell_q=1}^{L_q+1} \sum_{\ell_{q'}=1}^{L_{q'}+1} \sum_{i(s,\ell_q,\ell_{q'})} \log \Big[\Phi^{\ell_q,\ell_{q'}}_{i(s,\ell_q,\ell_{q'})qq'} - \Phi^{\ell_q-1,\ell_{q'}}_{i(s,\ell_q,\ell_{q'})qq'} \\ & - \Phi^{\ell_q,\ell_{q'}-1}_{i(s,\ell_q,\ell_{q'})qq'} + \Phi^{\ell_q-1,\ell_{q'}-1}_{i(s,\ell_q,\ell_{q'})qq'} \Big],\end{aligned} \tag{4.15}$$

where $\Phi^{\ell_q,\ell_{q'}}_{i(s,\ell_q,\ell_{q'})qq'}$ is the standard bivariate normal distribution function with correlation $r_{qq'}$ evaluated at $(\nu^{\ell_q}_{i(s,\ell_q,\ell_{q'})q}, \nu^{\ell_{q'}}_{i(s,\ell_q,\ell_{q'})q'})$. Expression (4.15) is then maximized with respect to $\boldsymbol{\theta}$ and its maximizer $\widehat{\boldsymbol{\theta}}^{PL}$ is called the *maximum pairwise likelihood* (MPL) estimator. Note that $\widehat{\boldsymbol{\theta}}_1^{PL} = \widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2^{PL} = \widehat{\boldsymbol{\theta}}_2$. Note as well that $\widehat{\boldsymbol{\theta}}_3^{PL}$ is a solution of the *pairwise score* equation $\mathbf{s}_{PL}(\boldsymbol{\theta}_3) = \mathbf{0}$, where $\mathbf{s}_{PL}(\boldsymbol{\theta}_3) = \partial \ell_3^p(\boldsymbol{\theta}_3)/\partial \boldsymbol{\theta}_3$ is the *pairwise score vector*. The elements of $\mathbf{s}_{PL}(\boldsymbol{\theta}_3)$ are given in the following lemma.

Lemma 4.3 *The elements of the pairwise score vector* $\mathbf{s}_{PL}(\boldsymbol{\theta}_3)$ *are given by the following:*

$$\frac{\partial \ell_3^p(\boldsymbol{\theta})}{\partial \gamma_q^{\ell_q}} = \sum_{s=1}^{S} \sum_{q',q'<q} \sum_{\ell_{q'}=1}^{L_{q'}+1} \sum_{i(s,\ell_q,\ell_{q'})} B^{-1}_{i(s,\ell_q,\ell_{q'})qq'} \phi^{\ell_q}_{i(s,\ell_q,\ell_{q'})q} \times \left[\Phi(\overline{\nu}^{\ell_{q'},\ell_q}_{i(s,\ell_q,\ell_{q'})qq'}) - \Phi(\overline{\nu}^{\ell_{q'}-1,\ell_q}_{i(s,\ell_q,\ell_{q'})qq'})\right],$$

$$\frac{\partial \ell_3^p(\boldsymbol{\theta})}{\partial r_{qq'}} = \sum_{s=1}^{S} \sum_{\ell_q=1}^{L_q+1} \sum_{\ell_{q'}=1}^{L_{q'}+1} \sum_{i(s,\ell_q,\ell_{q'})} B^{-1}_{i(s,\ell_q,\ell_{q'})qq'} \left[\phi^{\ell_q,\ell_{q'}}_{i(s,\ell_q,\ell_{q'})qq'} - \phi^{\ell_q-1,\ell_{q'}}_{i(s,\ell_q,\ell_{q'})qq'} - \phi^{\ell_q,\ell_{q'}-1}_{i(s,\ell_q,\ell_{q'})qq'} + \phi^{\ell_q-1,\ell_{q'}-1}_{i(s,\ell_q,\ell_{q'})qq'}\right],$$

$$\frac{\partial \ell_3^p(\boldsymbol{\theta})}{\partial \beta_q} = \sum_{s=1}^{S} \sum_{q',q'<q} \sum_{\ell_q=1}^{L_q+1} \sum_{\ell_{q'}=1}^{L_{q'}+1} \sum_{i(s,\ell_q,\ell_{q'})} B^{-1}_{i(s,\ell_q,\ell_{q'})qq'} \mathbf{y}_{i(s,\ell_q,\ell_{q'})} \times \left[\phi^{\ell_q}_{i(s,\ell_q,\ell_{q'})q} \Phi(\overline{\nu}^{\ell_{q'},\ell_q}_{i(s,\ell_q,\ell_{q'})q'q}) - \phi^{\ell_q-1}_{i(s,\ell_q,\ell_{q'})q} \Phi(\overline{\nu}^{\ell_{q'},\ell_q-1}_{i(s,\ell_q,\ell_{q'})q'q}) - \phi^{\ell_q}_{i(s,\ell_q,\ell_{q'})q} \Phi(\overline{\nu}^{\ell_{q'}-1,\ell_q}_{i(s,\ell_q,\ell_{q'})q'q}) + \phi^{\ell_q-1}_{i(s,\ell_q,\ell_{q'})q} \Phi(\overline{\nu}^{\ell_{q'}-1,\ell_q-1}_{i(s,\ell_q,\ell_{q'})q'q})\right],$$

$$\frac{\partial \ell_3^p(\boldsymbol{\theta})}{\partial \tau_{sq}} = \sum_{q',q'<q} \sum_{\ell_q=1}^{L_q+1} \sum_{\ell_{q'}=1}^{L_{q'}+1} \sum_{i(s,\ell_q,\ell_{q'})} B^{-1}_{i(s,\ell_q,\ell_{q'})qq'} \left[\phi^{\ell_q}_{i(s,\ell_q,\ell_{q'})q} \Phi(\overline{\nu}^{\ell_{q'}-1,\ell_q}_{i(s,\ell_q,\ell_{q'})q'q}) + \phi^{\ell_q-1}_{i(s,\ell_q,\ell_{q'})q} \Phi(\overline{\nu}^{\ell_{q'},\ell_q-1}_{i(s,\ell_q,\ell_{q'})q'q}) - \phi^{\ell_q}_{i(s,\ell_q,\ell_{q'})q} \Phi(\overline{\nu}^{\ell_{q'},\ell_q}_{i(s,\ell_q,\ell_{q'})q'q}) - \phi^{\ell_q-1}_{i(s,\ell_q,\ell_{q'})q} \Phi(\overline{\nu}^{\ell_{q'}-1,\ell_q-1}_{i(s,\ell_q,\ell_{q'})q'q})\right],$$

where $B_{i(s,\ell_q,\ell_{q'})qq'} = \Phi^{\ell_q,\ell_{q'}}_{i(s,\ell_q,\ell_{q'})qq'} - \Phi^{\ell_q-1,\ell_{q'}}_{i(s,\ell_q,\ell_{q'})qq'} - \Phi^{\ell_q,\ell_{q'}-1}_{i(s,\ell_q,\ell_{q'})qq'} + \Phi^{\ell_q-1,\ell_{q'}-1}_{i(s,\ell_q,\ell_{q'})qq'}$ *and* $\overline{\nu}^{\ell_{q'},\ell_q}_{i(s,\ell_q,\ell_{q'})q'q} = (\nu^{\ell_{q'}}_{i(s,\ell_q,\ell_{q'})q'} - r_{qq'}\nu^{\ell_q}_{i(s)q})/\sqrt{1-r^2_{qq'}}$.

Proof. The proof is straightforward. See also § 3.4 and 3.7.

□

The pairwise score equation may be solved iteratively using the Fletcher-Powell algorithm or a modified Fisher scoring method presented in § 3.4 (Kuk

and Nott, 2000). The same initial estimates given in (4.12) may be used to start the algorithm. Note that the pairwise approach results in a considerable reduction in computing time as only univariate and bivariate normal distributions are considered.

Unlike PML methods, the maximum pairwise likelihood method outlined above estimates the ordinal data parameters simultaneously, which results in a single set of parameter estimates. This avoids the problem of combining several sets of estimates, which is done in the PML methods. Note, however, that MPL estimation is very similar to the pairwise PML method of Bedrick *et al.* (2000) and Lapidus (1998) in that the probit model is similarly broken down into its bivariate sub-models by the former from which the pseudo-likelihood function is constructed. For additional details as well as its performance in terms of efficiency and bias, see Chapter 3.

4.5 Asymptotic Distributions of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}^{PL}$

Standard large-sample results on maximum likelihood estimation (Lehmann, 2000; Rao, 1973) and generalized score equations (Boos, 1992) are employed below to derive the asymptotic distributions of $\widehat{\boldsymbol{\theta}}$ in § 4.3 and $\widehat{\boldsymbol{\theta}}^{PL}$ in § 4.4.

Theorem 4.4 *The* MLE $\widehat{\boldsymbol{\theta}}$ *of* $\boldsymbol{\theta}$ *is consistent and satisfies*

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \xrightarrow{\mathcal{L}} \mathcal{N}_P\left(\mathbf{0}, \frac{1}{N}\mathcal{I}_P^{-1}(\boldsymbol{\theta})\right),$$

as $N \to \infty$, *with*

$$\mathcal{I}_P^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \mathcal{I}_{P_1}^{-1}(\boldsymbol{\theta}_1) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{I}_{P_2}^{-1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathcal{I}_{P_3}^{-1}(\boldsymbol{\theta}) \end{pmatrix}, \tag{4.16}$$

where $\mathcal{I}_{P_1}^{-1}(\boldsymbol{\theta}_1) = \text{diag}(\pi_1, \cdots, \pi_{S-1}) - \boldsymbol{\theta}_1\boldsymbol{\theta}_1^\top$, $\mathcal{I}_{P_2}^{-1}(\boldsymbol{\theta}_2)$ *is given by*

$$\frac{1}{N}\mathcal{I}_{P_2}^{-1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \begin{pmatrix} \mathrm{E}_{\boldsymbol{\theta}}(\mathbf{I}^*) \otimes \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Upsilon} \end{pmatrix}, \tag{4.17}$$

$\boldsymbol{\Upsilon}_{(P_2-CS)\times(P_2-CS)}$ *contains the asymptotic variances and covariances of the unique elements of* $\mathbf{S}$,

$$\mathrm{E}_{\boldsymbol{\theta}}(\mathbf{I}^*) = \text{diag}\left(\cdots, \frac{1-(N+1)\pi_s(1-\pi_s)^{N+1}}{(N+1)\pi_s[1-\pi_s^N-(1-\pi_s)^N]}, \cdots\right),$$

$N\mathcal{I}_{P_3}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[-\mathbf{H}(\boldsymbol{\theta}_3)]$ *with* $\mathbf{H}(\boldsymbol{\theta}_3)$ *as defined in § 4.3,* $P_1 = S-1$, $P_2 = CS + C(C+1)/2$, *and* $P_3 = P - P_1 - P_2$.

Proof. Consistency and normality of $\widehat{\boldsymbol{\theta}}$ follow from Theorem 7.5.2 of Lehmann (2000, p. 501), since the general mixed-data model can be easily shown to satisfy regularity conditions $(M1)-(M5)$, $(M6)''$, and $(M7)-(M8)$ in Lehmann (2000, pp. 499-501). Expression (4.16) is obtained by direct calculation.

The elements of $\mathrm{E}_{\boldsymbol{\theta}}[-\mathbf{H}(\boldsymbol{\theta}_3)]$ can be obtained from Lemma 4.1 in the case $Q = 1$, and from Lemma 4.2 in the general case $Q \geq 2$ by noting that $\mathrm{E}_{\boldsymbol{\theta}}[-\mathbf{H}(\boldsymbol{\theta}_3)] = \mathrm{E}_{\boldsymbol{\theta}}[\mathbf{s}(\boldsymbol{\theta}_3)\mathbf{s}^\top(\boldsymbol{\theta}_3)]$ (Mardia *et al.*, 1979, p. 98). The elements of $\mathrm{E}_{\boldsymbol{\theta}}(\mathbf{I}^*)$ are obtained by noting that $\mathrm{E}_{\boldsymbol{\theta}}(1/n_s) = 1/[\pi_s(N+1)]$ $\forall s$. Those of $\mathbf{A}$ in (4.17) are given in Theorem 3.2 in Chapter 3.

□

Remark 4.5.1 *Since* $\ell_3(\boldsymbol{\theta}_3)$ *is a conditional log-likelihood given* $n_1, \cdots, n_S$ *and* $\{\cdots, \mathbf{y}_{1(s)}, \cdots, \mathbf{y}_{n_s(s)}, \cdots\}$, *note that the information matrix* $\mathcal{I}_{P_3}(\boldsymbol{\theta})$ *associated with* $\boldsymbol{\theta}_3$ *is obtained by taking the expectation of* $\mathbf{H}(\boldsymbol{\theta}_3)$ *with respect to*

$[\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i]$. *This results in an expression that depends not only on* $\boldsymbol{\theta}_3$ *but on* $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ *as well.*

Theorem 4.5 extends Theorem 3.2 to the general mixed-data model.

Theorem 4.5 *Under the regularity conditions* A1 – A6 *given in Chapter 3, the* MPL *estimator* $\widehat{\boldsymbol{\theta}}^{PL}$ *of* $\boldsymbol{\theta}$ *is consistent and satisfies*

$$\widehat{\boldsymbol{\theta}}^{PL} - \boldsymbol{\theta} \xrightarrow{\mathcal{L}} \mathcal{N}_P(\mathbf{0}, \mathbf{V}),$$

as $N \to \infty$, *where*

$$\mathbf{V} = \begin{pmatrix} \frac{1}{N}\mathcal{I}_{P_1}^{-1}(\boldsymbol{\theta}_1) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{N}\mathcal{I}_{P_2}^{-1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_{P_3}^{-1}\mathbf{K}_{P_3}\mathbf{J}_{P_3}^{-1} \end{pmatrix},$$

$\mathbf{J}_{P_3} \equiv \mathbf{J}_{P_3}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[-\partial \mathbf{s}_{PL}(\boldsymbol{\theta}_3)/\partial \boldsymbol{\theta}_3^\top]$, *and* $\mathbf{K}_{P_3} \equiv \mathbf{K}_{P_3}(\boldsymbol{\theta})$ *is defined as*

$$\begin{aligned} \mathbf{K}_{P_3} &= \sum_{s=1}^{S} \sum_{i(s)} \mathrm{E}_{\boldsymbol{\theta}} \left[\left(\sum_{q<q'} \frac{\partial \ell^{p}_{i(s)qq'}(\boldsymbol{\theta}_3)}{\partial \boldsymbol{\theta}_3} \right) \right. \\ &\quad \left. \times \left(\sum_{q<q'} \frac{\partial \ell^{p}_{i(s)qq'}(\boldsymbol{\theta}_3)}{\partial \boldsymbol{\theta}_3} \right)^\top \right], \end{aligned}$$

with $\ell^{p}_{i(s)qq'}(\boldsymbol{\theta}_3)$ *is the pairwise log-likelihood contribution of* $\{Z_q, Z_{q'}\}$ *from the* i*th observation in the* s*th state.*

Proof. The proof follows along the same lines as that of Theorem 3.2 and similarly relies on results in Guyon (1995), Boos (1992), and Crowder (1986). To show asymptotic normality, define the *pairwise information matrix* $\mathbf{J}_P \equiv \mathbf{J}_P(\boldsymbol{\theta})$ as follows:

$$\begin{aligned} \mathbf{J}_P &= \mathrm{E}_{\boldsymbol{\theta}} \left[-\frac{\partial^2 \ell_1(\boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right] + \mathrm{E}_{\boldsymbol{\theta}} \left[-\frac{\partial^2 \ell_2(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right] + \mathrm{E}_{\boldsymbol{\theta}} \left[-\frac{\partial^2 \ell^{p}_3(\boldsymbol{\theta}_3)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right] \\ &= \begin{pmatrix} N\mathcal{I}_{P_1}(\boldsymbol{\theta}_1) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & N\mathcal{I}_{P_2}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_{P_3} \end{pmatrix}, \end{aligned}$$

where the last equality follows from the fact that $\widehat{\boldsymbol{\theta}}_1^{PL} = \widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2^{PL} = \widehat{\boldsymbol{\theta}}_2$, and $\mathbf{J}_{P_3} = \mathrm{E}_{\boldsymbol{\theta}}[-\partial^2 \ell_3^p(\boldsymbol{\theta}_3)/\partial\boldsymbol{\theta}_3\partial\boldsymbol{\theta}_3^\top]$. Also, define $\mathbf{K}_P \equiv \mathbf{K}_P(\boldsymbol{\theta})$ as

$$\begin{aligned}\mathbf{K}_P &= \sum_{s=1}^{S}\sum_{i(s)} \mathrm{E}_{\boldsymbol{\theta}}\left[\left(\sum_{q<q'} \frac{\partial \ell^p_{i(s)qq'}(\boldsymbol{\theta}_3)}{\partial \boldsymbol{\theta}_3}\right)\left(\sum_{q<q'} \frac{\partial \ell^p_{i(s)qq'}(\boldsymbol{\theta}_3)}{\partial \boldsymbol{\theta}_3}\right)^\top\right] \\ &= \begin{pmatrix} N\mathcal{I}_{P_1}(\boldsymbol{\theta}_1) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & N\mathcal{I}_{P_2}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{K}_{P_3} \end{pmatrix},\end{aligned}$$

where the last equality follows from the fact that $\mathbf{K}_{P_1}(\boldsymbol{\theta}_1) = N\mathcal{I}_{P_1}(\boldsymbol{\theta}_1)$ and $\mathbf{K}_{P_2}(\boldsymbol{\theta}_2) = N\mathcal{I}_{P_2}(\boldsymbol{\theta}_2)$ (Mardia *et al.*, 1979, p. 98).

Observing that $\mathrm{E}_{\boldsymbol{\theta}}[\mathbf{s}_{PL}(\boldsymbol{\theta}_3)] = \mathbf{0}$ and assuming $A1 - A6$ hold, it can be shown (Boos, 1992) that

$$\widehat{\boldsymbol{\theta}}^{PL} - \boldsymbol{\theta} = \mathbf{J}_P^{-1}\begin{pmatrix} \frac{\partial}{\partial\boldsymbol{\theta}_1}\ell_1(\boldsymbol{\theta}_1) \\ \frac{\partial}{\partial\boldsymbol{\theta}_2}\ell_2(\boldsymbol{\theta}_2) \\ \mathbf{s}_{PL}(\boldsymbol{\theta}_3) \end{pmatrix} + o_p(N),$$

where $o_p(N) \xrightarrow{p} 0$ as $N \to \infty$. Thus, $\widehat{\boldsymbol{\theta}}^{PL} - \boldsymbol{\theta}$ is asymptotically multivariate normal with asymptotic mean $\mathbf{0}$ and asymptotic covariance matrix

$$\begin{aligned}\mathbf{V} &= \mathbf{J}_P^{-1}\mathbf{K}_P\mathbf{J}_P^{-1} \\ &= \begin{pmatrix} \frac{1}{N}\mathcal{I}_{P_1}^{-1}(\boldsymbol{\theta}_1) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{N}\mathcal{I}_{P_2}^{-1}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_{P_3}^{-1}\mathbf{K}_{P_3}\mathbf{J}_{P_3}^{-1} \end{pmatrix}.\end{aligned}$$

□

Theorems 4.4 and 4.5 are used in the subsequent section to construct asymptotic tests of hypotheses concerning $\boldsymbol{\theta}$.

4.6 Statistical Inference

Typical inferential questions of interest involve comparisons of the different state means $\boldsymbol{\mu}_s$ of the continuous variables, the level- and state-specific effects

τ_{sq}, the regression coefficients $\boldsymbol{\beta}_q$, and the polychoric correlations $r_{qq'}$. For example, the following hypotheses may be of particular interest:

$$
\begin{aligned}
H_1 &: \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_S, \\
H_2 &: \tau_{11} = \cdots = \tau_{S-1,q} = 0 \quad (q = 1, \cdots, Q), \\
H_3 &: \tau_{s1} = \cdots = \tau_{sQ} = 0 \quad (s = 1, \cdots, S-1), \\
H_4 &: \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_Q = \mathbf{0}, \\
H_5 &: \tau_{qq'} = 0 \quad (q < q'; q, q' = 1, \cdots, Q).
\end{aligned}
$$

Hypothesis H_1 tests the independence of the nominal and continuous variables. Hypothesis H_2 tests the absence of level-specific effect while H_3 corresponds to that for state-specific effect. Hypothesis H_4 is equivalent to $H_2 \cap H_3$, the test of independence of the continuous and ordinal variables. Finally, H_5 concerns the independence of the ordinal variables.

Note that hypothesis H_1 is easily tested using the statistics given by Olkin and Tate (1961, Theorems 4.1 and 5.1) and Morales *et al.* (1998).

To construct tests of hypotheses such as those above, likelihood ratio, Wald and score test statistics can be constructed based on usual asymptotic theory. For the MLE $\widehat{\boldsymbol{\theta}}$, $\mathcal{I}_P(\boldsymbol{\theta})$ must be estimated by a consistent estimator $\widehat{\mathcal{I}}_P(\boldsymbol{\theta})$, which may be obtained by replacing $\mathcal{I}_{P_1}(\boldsymbol{\theta}_1)$, $\mathcal{I}_{P_2}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, and $\mathcal{I}_{P_3}(\boldsymbol{\theta})$ with, respectively, $\mathcal{I}_{P_1}(\widehat{\boldsymbol{\theta}}_1)$, $\mathcal{I}_{P_2}(\widehat{\boldsymbol{\theta}}_1, \widehat{\boldsymbol{\theta}}_2)$, and $\mathbf{s}(\widehat{\boldsymbol{\theta}}_3)\mathbf{s}^\top(\widehat{\boldsymbol{\theta}}_3) = \mathbf{s}(\boldsymbol{\theta}_3)\mathbf{s}^\top(\boldsymbol{\theta}_3)|_{\boldsymbol{\theta}_3 = \widehat{\boldsymbol{\theta}}_3}$, the observed Fisher information matrix evaluated at $\widehat{\boldsymbol{\theta}}_3$. Note that $\mathbf{s}(\widehat{\boldsymbol{\theta}}_3)\mathbf{s}^\top(\widehat{\boldsymbol{\theta}}_3)$ is independent of $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$.

If $\widehat{\boldsymbol{\theta}}^{PL}$ is used instead, the so-called *information sandwich* estimate may

be used for $\mathbf{V}$. Note that in the present case, the first two block matrices along the diagonal of $\mathbf{V}$ are estimated as in the MLE case, while $(\widehat{\mathbf{J}}_{P_3}^{PL})^{-1}\widehat{\mathbf{K}}_{P_3}(\widehat{\mathbf{J}}_{P_3}^{PL})^{-1}$ is used for the third block diagonal, where $\widehat{\mathbf{J}}_{P_3}^{PL}$ is as defined in Chapter 3, and

$$\widehat{\mathbf{K}}_{P_3} = \sum_{s=1}^{S}\sum_{i(s)}\left(\sum_{q<q'}\frac{\partial \ell^{p}_{i(s)qq'}(\boldsymbol{\theta}_3)}{\partial \boldsymbol{\theta}_3}\right)\left(\sum_{q<q'}\frac{\partial \ell^{p}_{i(s)qq'}(\boldsymbol{\theta}_3)}{\partial \boldsymbol{\theta}_3}\right)^{\top},$$

again evaluated at $\widehat{\boldsymbol{\theta}}_3^{PL}$.

Theorems 4.4 and 4.5 may now be used to construct generalized Wald and likelihood ratio tests of hypotheses concerning $\boldsymbol{\theta}$. For instance, suppose it is desired to test $H : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_Q$, the hypothesis of uniform polyserial correlation structure, using the MPL estimate $\widehat{\boldsymbol{\theta}}^{PL}$. Note that H is equivalent to the hypothesis $H' : \mathbf{C}\boldsymbol{\theta} = \mathbf{0}$, where $\mathbf{C} = (\mathbf{0}, \mathbf{C}_{\boldsymbol{\beta}}, \mathbf{0})$, and $\mathbf{C}_{\boldsymbol{\beta}}$ is a $C(Q-1) \times CQ$ matrix defined by

$$\mathbf{C}_{\boldsymbol{\beta}} = \begin{pmatrix} \mathbf{I}_C & \mathbf{0} & \cdots & \mathbf{0} & -\mathbf{I}_C \\ \mathbf{0} & \mathbf{I}_C & \cdots & \mathbf{0} & -\mathbf{I}_C \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_C & -\mathbf{I}_C \end{pmatrix}.$$

Since $\text{rank}(\mathbf{C}) = C(Q-1)$, a Wald-type large-sample χ^2 statistic can then be constructed as

$$X_W^2 = (\mathbf{C}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{PL})^{\top}\left(\mathbf{C}\widehat{\mathbf{V}}^{PL}\mathbf{C}^{\top}\right)^{-1}(\mathbf{C}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^{PL}),$$

and by Theorem 4.5, $X_W^2 \xrightarrow{\mathcal{L}} \chi^2_{C(Q-1)}$ under H. Here, $\widehat{\boldsymbol{\beta}}^{PL}$ and $\widehat{\mathbf{V}}^{PL}$ are obtained using the MPL estimates.

4.7 Appendicitis Data Example

In this section, real data are considered to illustrate the general mixed-data model. The data come from Koepsel *et al.* (1981) (also found on pp. 680-683

of Fisher and Van Bell, 1993) and concern the occurrence and non-occurrence of perforation of the appendix. Data from a total of 181 surgery patients are included in the analysis, and three variables are considered. The same data were analyzed by Nakanishi (1996) in the context of variable selection in mixed-data discriminant analysis. For the purposes of this example, only those subjects with waiting times to surgery exceeding 0 but not exceeding 60 hours were included in the analysis. In addition, the waiting times to surgery were transformed using their natural logarithms. Normal probability plots of the transformed waiting times indicate that the assumption of normality is satisfied.

In what follows, the variable X_3 as defined in Fisher and Van Bell (1993, p. 680) is transformed into an ordinal variable Z with 2 levels (long or short duration). The states of $\mathbf{x}^\top = (X_1, X_2)$ correspond with the patient's perforation status, with $\mathbf{x} = \mathbf{x}_{(2)}$ if perforation is present and $\mathbf{x} = \mathbf{x}_{(1)}$ otherwise. The following variables are included:

$$
\begin{aligned}
Y &:= \text{time in hours from physician contact to surgery,} \\
Z &:= \text{duration of symptoms prior to physician contact} \\
&= \begin{cases} 2 & \text{no. of hours} > 24 \\ 1 & \text{otherwise} \end{cases}
\end{aligned}
$$

A general mixed-data model with $C = L = Q = 1$ and $S = 2$ is fit to these data. The parameter is then $\boldsymbol{\theta}^\top = (\pi, \boldsymbol{\mu}^\top, \sigma^2, \gamma, \beta, \tau)$, where $\boldsymbol{\mu}^\top = (\mu_1, \mu_2)$ with μ_s the sth state mean of Y, γ is the standardized cutpoint α for the latent variable Y^* underlying Z, and τ is the effect of state 1 on Z relative to that of state 2. Note that $\gamma = \alpha/\sqrt{1-\rho^2} - (\mu_2^*/\sqrt{1-\rho^2} - \beta\mu_2)$, $\beta = \rho/(\sigma\sqrt{1-\rho^2})$,

and $\tau = \xi^*/\sqrt{1-\rho^2} - \beta\xi$, where $\xi = \mu_1 - \mu_2$, $\xi^* = \mu_1^* - \mu_2^*$, with $\mu_s^* = \mathrm{E}(Y^* \mid \mathbf{x} = \mathbf{x}_{(s)})$, for $s = 1, 2$.

The MLEs (which are also the MPLEs in this case) of the parameters were calculated using S-PLUS, and are presented in Table 4.2, along with their corresponding standard errors. From the estimates in Table 4.2, the estimated point-biserial correlation between perforation status and the logarithm of waiting time to surgery is found to be -0.1374, indicating a weak negative association. This agrees with the conclusions of Koepsel *et al.* (1981), who found that perforation status is only strongly related to the length of the pre-admission phase, the period prior to physician contact. In addition, the correlation between duration of symptoms and logarithm of waiting time to surgery is estimated as -0.2236.

Wald test statistics for testing $H_\tau : \tau = 0$ and $H_\beta : \beta = 0$ yielded $X_\tau^2 = 44.76$ and $X_\beta^2 = 9.07$, respectively. Upon comparison with the critical value $\chi^2_{1,0.05} = 3.8414$, it can be concluded that τ (p-value<0.05) and β (p-value=0.0026) are both significantly different from 0. The conclusion regarding the parameter τ confirms Koepsel *et al.*'s (1981) finding regarding the positive association between perforation status and duration of symptoms.

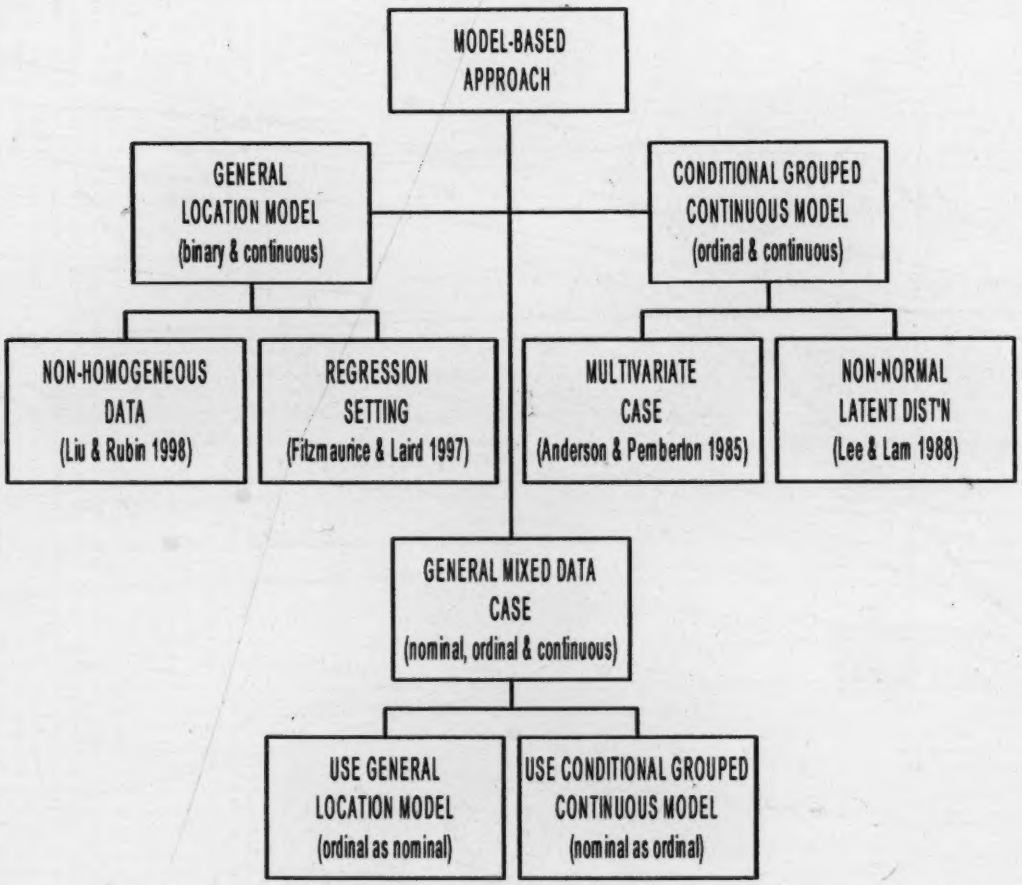
Modelling the data via the general mixed-data model makes these detailed information possible while maintaining the levels of measurement of the variables.

4.8 Discussion

In this chapter, a general model for multivariate data with mixtures of nominal, ordinal and continuous variables called the *general mixed-data model* was proposed. The approach adopted in developing the model is motivated by the need to account for the various levels of measurement, and hence, the different types of information, in the data, which many conventional approaches fail to incorporate in the analysis. The general mixed-data model in Definition 4.1 includes as special cases the general location model of Olkin and Tate (1961) and the mixed-data models studied by Bedrick *et al.* (2000), Lapidus (1998), Poon and Lee (1987; 1986), and Anderson and Pemberton (1985).

A full likelihood-based approach that yields maximum likelihood estimates of the model parameters was outlined, and algorithms to implement it were provided. The methods generalize earlier works of Bedrick *et al.* (2000), Lapidus (1998), and Poon and Lee (1987; 1986). An alternative based on the pairwise likelihood approach was also presented. Statistical inference for various hypotheses on comparisons of the means, polychoric and polyserial parameters among the variables across and within states based on both approaches was also discussed.

Figure 4.1: *Taxonomy of Models and Analytical Approaches in Mixed Multivariate Data Analysis.*


MODEL-BASED APPROACH
GENERAL LOCATION MODEL (binary & continuous)
CONDITIONAL GROUPED CONTINUOUS MODEL (ordinal & continuous)
NON-HOMOGENEOUS DATA (Liu & Rubin 1998)
REGRESSION SETTING (Fitzmaurice & Laird 1997)
MULTIVARIATE CASE (Anderson & Pemberton 1985)
NON-NORMAL LATENT DIST'N (Lee & Lam 1988)
GENERAL MIXED DATA CASE (nominal, ordinal & continuous)
USE GENERAL LOCATION MODEL (ordinal as nominal)
USE CONDITIONAL GROUPED CONTINUOUS MODEL (nominal as ordinal)

NOTE: *Displayed is a tree chart of the various models and approaches adopted in mixed-data analysis, including their extensions and their references.*

Figure 4.2: *Two Levels of Data Layout for the General Mixed-Data Model with* $Q = S = 2, L_1 = L_2 = 1$ *and* $C \geq 1$

Layout at the State Level

State 1	State 2
$\{\mathbf{y}_{1(1)}, \mathbf{z}_{1(1)}\}$	$\{\mathbf{y}_{1(2)}, \mathbf{z}_{1(2)}\}$
$\vdots$	$\vdots$
$\{\mathbf{y}_{n_1(1)}, \mathbf{z}_{n_1(1)}\}$	$\{\mathbf{y}_{n_2(2)}, \mathbf{z}_{n_2(2)}\}$

Layout at the Level of the Ordinal Vector

	State 1	State 2
$\boldsymbol{\ell}^\top = (1,1):$	$\mathbf{y}_{1(1,1,1)}, \cdots, \mathbf{y}_{n_{111}(1,1,1)}$	$\mathbf{y}_{1(2,1,1)}, \cdots, \mathbf{y}_{n_{211}(2,1,1)}$
$\boldsymbol{\ell}^\top = (1,2):$	$\mathbf{y}_{1(1,1,2)}, \cdots, \mathbf{y}_{n_{112}(1,1,2)}$	$\mathbf{y}_{1(2,1,2)}, \cdots, \mathbf{y}_{n_{212}(2,1,2)}$
$\boldsymbol{\ell}^\top = (2,1):$	$\mathbf{y}_{1(1,2,1)}, \cdots, \mathbf{y}_{n_{121}(1,2,1)}$	$\mathbf{y}_{1(2,2,1)}, \cdots, \mathbf{y}_{n_{221}(2,2,1)}$
$\boldsymbol{\ell}^\top = (2,2):$	$\mathbf{y}_{1(1,2,2)}, \cdots, \mathbf{y}_{n_{122}(1,2,2)}$	$\mathbf{y}_{1(2,2,2)}, \cdots, \mathbf{y}_{n_{222}(2,2,2)}$

NOTE: *The vector* $\boldsymbol{\ell}$ *is a realized value of the ordinal vector* $\mathbf{z}^\top_{i(s)} = (Z_{i(s)1}, Z_{i(s)2})$. *Note that* $n_s = \sum_{\ell_1=1}^{2} \sum_{\ell_2=1}^{2} n_{s\ell_1\ell_2}$, $s = 1, 2$, *and* $n_1 + n_2 = N$. *Also, note, for instance, that* $\{\mathbf{y}_{1(1,1,1)}, \cdots, \mathbf{y}_{n_{111}(1,1,1)}, \cdots, \mathbf{y}_{1(1,2,2)}, \cdots, \mathbf{y}_{n_{122}(1,2,2)}\}$ *is simply a relabelling of* $\{\mathbf{y}_{1(1)}, \cdots, \mathbf{y}_{n_1(1)}\}$.

Table 4.1: *Three-Dimensional Array for the Appendicitis Data (Koepsel et al., 1981).*

	Perforation		
Duration	Yes	No	Total
> 24 hrs.	28	40	68
≤ 24 hrs.	10	103	113
Total	38	143	181

NOTE: *Shown are the numbers of surgery patients classified according to population (male or female), perforation state (s=1 if perforation is present and s=2 otherwise), and duration (Z=2 if duration exceeds 24 hrs. and Z=1 otherwise). The actual values of the time Y from diagnosis to surgery are found in Fisher and Van Bell (1993, p. 680).*

Table 4.2: *Maximum Likelihood Estimates of Parameters of General Mixed-Data Model for the Appendicitis Data.*

Parameter	Estimate	Standard Error
$\widehat{p}$	0.2099	0.0302
$\widehat{\mu}_1$	1.2032	0.1544
$\widehat{\mu}_2$	1.5622	0.0796
$\widehat{\gamma}$	0.9585	0.1454
$\widehat{\beta}$	0.2404	0.0798
$\widehat{\tau}$	1.2912	0.1929
$\widehat{\sigma}$	0.9542	0.0954

NOTE: *Shown are the maximum likelihood estimates of the general mixed-data model parameters for the appendicitis data.*

Chapter 5

A Generalization of Mahalanobis Distance to Mixed Qualitative and Quantitative Data

5.1 Introduction

The estimation of a statistical distance between populations arises in many multivariate analysis techniques. In cluster analysis, for example, a *dissimilarity measure*, defined by a distance metric, is needed to evaluate the proximity of two observations (Seber, 1984, pp. 351-355). The same scenario may be found in some discrimination problems (Dillon and Goldstein, 1978). Whereas distance measures for use with continuous data are well developed (Seber, 1984), those for mixed discrete and continuous data are less so because of the lack of a standard model for such data.

Krzanowski (1984; 1983) was the first to consider the development of mixed data distances. After applying *Matusita's distance* (Matusita, 1956) to the general location model, Krzanowski (1983) derived a distance measure

between two groups based on mixed nominal and continuous data. The exact and asymptotic distributions of its sample estimates under the null hypothesis of non-distinct groups were later studied by Krzanowski (1984) and Bar-Hen and Daudin (1998), respectively. Bar-Hen and Daudin (1995), in contrast, applied the *Kullback-Leibler divergence* (Kullback, 1968, pp. 6-7) to the general location model and obtained a distance that specializes to the Mahalanobis distance in the absence of nominal variables. Krusińska (1987) proposed a weighted Mahalanobis distance for mixed data as the weighted sum of the Mahalanobis distance for continuous variables and a Mahalanobis-type distance for discrete variables introduced by Kurczyński (1970). Krusińska and Liebhart (1988) later applied the weighted distance in outlier detection problems.

Besides the distances introduced by Bedrick *et al.* (2000), Bar-Hen and Daudin (1995), and Krzanowski (1983), no distance measure has yet been developed for mixed data with nominal, ordinal and continuous variables. Such a distance must account for not only the different levels of measurement in the variables but also the various types of associations among the variables.

The aim of this chapter is to develop a statistical distance that can be used for data consisting of a mixture of variable types. Specifically, the problem of generalizing the *Mahalanobis distance* (Mardia *et al.*, 1979, p. 31) to mixed data with nominal, ordinal and continuous variables is considered. The approach adopted in the chapter unifies previous work on the problem by Bedrick *et al.* (2000), Lapidus (1998), and Bar-Hen and Daudin (1995). The

latter extended the Mahalanobis distance to mixed nominal and continuous data via the general location model while the former used the grouped continuous model for mixed ordinal and continuous data and derived a Mahalanobis distance for the data.

The chapter is organized as follows. A general distance measure for mixed nominal, ordinal and continuous data is developed in § 5.2, and the asymptotic distribution of its MLE is obtained. In addition, large-sample tests of hypothesis concerning two mixed-variate populations are also derived. The finite-sample performance of these tests are investigated via simulations in § 5.4. A real-data example is presented in § 5.5 to illustrate the utility of the distance measure. Finally, the chapter concludes with a discussion in § 5.6.

5.2 A Generalized Mahalanobis Distance

In this section, a distance for mixed nominal, ordinal and continuous data as modelled by the general mixed-data model in Chapter 4, is derived. The distance includes as special cases previous generalizations of the Mahalanobis distance to mixed data proposed by Bedrick *et al.* (2000) and Bar-Hen and Daudin (1995).

Adopting the notations in § 2.4, suppose $(\mathbf{x}_g^\top, \mathbf{y}_g^\top, \mathbf{z}_g^\top)^\top$ is a random vector from the mixed-variate population $\mathcal{P}^{(g)}$ defined by the general mixed-data model with parameter $\boldsymbol{\theta}_g$ containing $\boldsymbol{\pi}_g$, $\boldsymbol{\mu}_g$ and $\boldsymbol{\tau}_g$, for $g = 1, \cdots, G$, and $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, vech($\boldsymbol{\Sigma}$), and vech($\mathbf{R}$). Note that this implies that the populations only differ

in their locations. As well, it is assumed that the reference states in each of the populations are the same with $\boldsymbol{\xi}_g^* = \boldsymbol{\xi}^*$ and $\boldsymbol{\xi}_g = \boldsymbol{\xi}$ $\forall g$. This approach is similar to that adopted earlier by Poon and Lee (1992) and Lee *et al.* (1989).

The following formal definition of the Kullback-Leibler divergence given by Kullback (1968, p. 6) is presented for later use.

Definition 5.1 *Let $\psi_{g'}$, $\psi_{g''}$ and λ be three probability measures absolutely continuous with respect to each other, and assume there exist generalized probability densities $f_{g'}$ and $f_{g''}$, the respective Radon-Nikodym derivatives of $\psi_{g'}$ and $\psi_{g''}$ with respect to λ. The divergence measure between $f_{g'}$ and $f_{g''}$ defined as*

$$\Delta_{g'g''} = \int [f_{g'}(\mathbf{w}) - f_{g''}(\mathbf{w})] \log \frac{f_{g'}(\mathbf{w})}{f_{g''}(\mathbf{w})} \mathrm{d}\lambda,$$

is called the Kullback-Leibler divergence.

Remark 5.2.1 *$\Delta_{g'g''}$ possesses all the properties of a distance except for the triangle inequality, and is therefore not considered a distance (Kullback, 1968, Chapter 2).*

Remark 5.2.2 *When $f_{g'}$ is $\mathcal{N}(\boldsymbol{\mu}_{g'}, \boldsymbol{\Sigma})$ and $f_{g''}$ is $\mathcal{N}(\boldsymbol{\mu}_{g''}, \boldsymbol{\Sigma})$, then $\Delta_{g'g''} = (\boldsymbol{\mu}_{g'} - \boldsymbol{\mu}_{g''})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{g'} - \boldsymbol{\mu}_{g''})$, the Mahalanobis distance between two multivariate normal populations. In this respect, $\Delta_{g'g''}$ can be considered as a generalization of the Mahalanobis distance.*

Remark 5.2.3 *Bar-Hen and Daudin (1995) used $\Delta_{g'g''}$ to generalize the Mahalanobis distance to mixed binary and continuous data modelled by the general location model, and derived the asymptotic distribution of its MLE.*

Theorem 5.1 below is obtained by applying Definition 5.1 to the general mixed-data models for $\mathcal{P}^{(g')}$ and $\mathcal{P}^{(g'')}$.

Theorem 5.1 *The Kullback-Leibler divergence between $\mathcal{P}^{(g')}$ and $\mathcal{P}^{(g'')}$ is given by*

$$\Delta_{g'g''} = \Delta^1_{g'g''} + \Delta^2_{g'g''} + \Delta^3_{g'g''}, \tag{5.1}$$

where

$$\begin{aligned}
\Delta^1_{g'g''} &= \sum_{s=1}^{S} (\pi_{g's} - \pi_{g''s}) \log \frac{\pi_{g's}}{\pi_{g''s}}, \\
\Delta^2_{g'g''} &= \sum_{s=1}^{S} \frac{\pi_{g's} + \pi_{g''s}}{2} \left(\boldsymbol{\mu}_{g's} - \boldsymbol{\mu}_{g''s}\right)^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}_{g's} - \boldsymbol{\mu}_{g''s}\right), \\
\Delta^3_{g'g''} &= \sum_{s=1}^{S-1} \frac{\pi_{g's} + \pi_{g''s}}{2} \left(\boldsymbol{\tau}_{g's} - \boldsymbol{\tau}_{g''s}\right)^\top \mathbf{R}^{-1} \left(\boldsymbol{\tau}_{g's} - \boldsymbol{\tau}_{g''s}\right).
\end{aligned}$$

Proof. Suppose $\mathbf{y}^*_g$ is the latent variable underlying $\mathbf{z}_g$, and that $(\mathbf{x}_g^\top, \mathbf{y}_g^\top, \mathbf{y}_g^{*\top})$ follows the $GLM(\boldsymbol{\pi}_g, (\boldsymbol{\mu}_g^\top, \boldsymbol{\mu}_g^{*\top})^\top, \boldsymbol{\Gamma})$, where $\boldsymbol{\mu}_g^{*\top} = (\boldsymbol{\mu}_{g1}^{*\top}, \cdots, \boldsymbol{\mu}_{gS}^{*\top})$ is the $QS \times 1$ stacked vector of state means of $\mathbf{y}^*_g$ and $\boldsymbol{\Gamma}$ is as defined in (4.2). Using results in Kullback (1968, Chapter 6) and Proposition 2.1 in Bar-Hen and Daudin (1995), it follows that

$$\begin{aligned}
\Delta_{g'g''} &= \sum_{s=1}^{S} (\pi_{g's} - \pi_{g''s}) \log \frac{\pi_{g's}}{\pi_{g''s}} \\
&\quad + \sum_{s=1}^{S} \frac{\pi_{g's} + \pi_{g''s}}{2} \begin{pmatrix} \boldsymbol{\mu}_{g's} - \boldsymbol{\mu}_{g''s} \\ \boldsymbol{\mu}^*_{g's} - \boldsymbol{\mu}^*_{g''s} \end{pmatrix}^\top \boldsymbol{\Gamma}^{-1} \begin{pmatrix} \boldsymbol{\mu}_{g's} - \boldsymbol{\mu}_{g''s} \\ \boldsymbol{\mu}^*_{g's} - \boldsymbol{\mu}^*_{g''s} \end{pmatrix}.
\end{aligned}$$

But by the decomposition of the Mahalanobis distance (Mardia *et al.*, 1979,

pp. 78-79),

$$\begin{pmatrix} \boldsymbol{\mu}_{g's} - \boldsymbol{\mu}_{g''s} \\ \boldsymbol{\mu}^*_{g's} - \boldsymbol{\mu}^*_{g''s} \end{pmatrix}^\top \boldsymbol{\Gamma}^{-1} \begin{pmatrix} \boldsymbol{\mu}_{g's} - \boldsymbol{\mu}_{g''s} \\ \boldsymbol{\mu}^*_{g's} - \boldsymbol{\mu}^*_{g''s} \end{pmatrix} = \left(\boldsymbol{\mu}_{g's} - \boldsymbol{\mu}_{g''s}\right)^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}_{g's} - \boldsymbol{\mu}_{g''s}\right) + \left(\boldsymbol{\mu}_{g's\cdot y} - \boldsymbol{\mu}_{g''s\cdot y}\right)^\top (\mathbf{DRD})^{-1} \times \left(\boldsymbol{\mu}_{g's\cdot y} - \boldsymbol{\mu}_{g''s\cdot y}\right),$$

where $\boldsymbol{\mu}_{gs\cdot y} = \boldsymbol{\mu}^*_{gs} - \mathbf{DB}\boldsymbol{\mu}_{gs}$, with $\mathbf{D}$ and $\mathbf{B}$ as defined in § 4.2, $g = g', g''$. Since $\boldsymbol{\mu}^*_{gs} = \boldsymbol{\xi}^* + \boldsymbol{\xi}^*_{gs}$ and $\boldsymbol{\mu}_{gs} = \boldsymbol{\xi} + \boldsymbol{\xi}_{gs}$ for $s \neq S$, it follows that $\boldsymbol{\mu}_{g's\cdot y} - \boldsymbol{\mu}_{g''s\cdot y} = \boldsymbol{\xi}^*_{g's} - \mathbf{DB}\boldsymbol{\xi}_{g's} - (\boldsymbol{\xi}^*_{g''s} - \mathbf{DB}\boldsymbol{\xi}_{g''s})$. Expression (5.1) is now immediate by noting from § 4.2 that $\boldsymbol{\tau}_{gs} = \mathbf{D}^{-1}\boldsymbol{\xi}^*_{gs} - \mathbf{B}\boldsymbol{\xi}_{gs}$ for $g = g', g''$.

□

Corollary 5.1.1 *If* $\boldsymbol{\mu}_{gs} = \boldsymbol{\mu}_g$ *and* $\boldsymbol{\mu}^*_{gs} = \boldsymbol{\mu}^*_g$ $\forall s$, *then*

$$\Delta_{g'g''} = \sum_{s=1}^{S} (\pi_{g's} - \pi_{g''s}) \log \frac{\pi_{g's}}{\pi_{g''s}} + \left(\boldsymbol{\mu}_{g'} - \boldsymbol{\mu}_{g''}\right)^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}_{g'} - \boldsymbol{\mu}_{g''}\right) + \left(\boldsymbol{\tau}_{g'} - \boldsymbol{\tau}_{g''}\right)^\top \mathbf{R}^{-1} \left(\boldsymbol{\tau}_{g'} - \boldsymbol{\tau}_{g''}\right).$$

Proof. By noting that $\boldsymbol{\mu}^*_{gs} = \boldsymbol{\mu}^*_g$ implies $\boldsymbol{\xi}^*_{gs} = \boldsymbol{\xi}^*_g$, it is clear that $\boldsymbol{\tau}_{gs} = \boldsymbol{\tau}_g$ $\forall s, g = g', g''$, and the proof is straightforward.

□

Further remarks concerning Theorem 5.1 and Corollary 5.1.1 are given below.

Remark 5.2.4 *With $Q = 0$, $\Delta_{g'g''} = \Delta^1_{g'g''} + \Delta^2_{g'g''}$ is the distance proposed by Bar-Hen and Daudin (1995) while with $S = 1$, $\Delta_{g'g''} = \Delta^2_{g'g''} + \Delta^3_{g'g''}$ corresponds to that by Bedrick* et al. *(2000) and Lapidus (1998). Thus, Theorem 5.1 generalizes these two previous Mahalanobis-type distances for mixed data.*

Remark 5.2.5 *$\Delta_{g'g''}$ can be considered an extension of the Mahalanobis distance since it reduces to it for $Q = 0, S = 1$. Note also that $\Delta_{g'g''} = \Delta_{g''g'}$ for any g', g''.*

Remark 5.2.6 *Corollary 5.1.1 states that when the nominal variables are independent of the continuous and ordinal variables, $\Delta_{g'g''}$ is simply the sum of the distances corresponding to each variable type.*

Remark 5.2.7 *As noted by Bedrick* et al. *(2000), the Mahalanobis distance*

$$\begin{pmatrix} \boldsymbol{\mu}_{g's} - \boldsymbol{\mu}_{g''s} \\ \boldsymbol{\mu}^*_{g's} - \boldsymbol{\mu}^*_{g''s} \end{pmatrix}^\top \boldsymbol{\Gamma}^{-1} \begin{pmatrix} \boldsymbol{\mu}_{g's} - \boldsymbol{\mu}_{g''s} \\ \boldsymbol{\mu}^*_{g's} - \boldsymbol{\mu}^*_{g''s} \end{pmatrix},$$

for fixed state s, remains estimable even though $\mathbf{y}^*_{gs}$ is unobservable.

Given random samples $(\mathbf{x}_{gi}^\top, \mathbf{y}_{gi}^\top, \mathbf{z}_{gi}^\top)^\top$, $i = 1, \cdots, n_g, g = 1, \cdots, G$, the MLE of $\Delta_{g'g''}$ is given by $\widehat{\Delta}_{g'g''} = \widehat{\Delta}^1_{g'g''} + \widehat{\Delta}^2_{g'g''} + \widehat{\Delta}^3_{g'g''}$, where

$$\begin{aligned}
\widehat{\Delta}^1_{g'g''} &= \sum_{s=1}^{S} (\widehat{\pi}_{g's} - \widehat{\pi}_{g''s}) \log \frac{\widehat{\pi}_{g's}}{\widehat{\pi}_{g''s}}, \\
\widehat{\Delta}^2_{g'g''} &= \sum_{s=1}^{S} \frac{\widehat{\pi}_{g's} + \widehat{\pi}_{g''s}}{2} \left(\widehat{\boldsymbol{\mu}}_{g's} - \widehat{\boldsymbol{\mu}}_{g''s}\right)^\top \widehat{\boldsymbol{\Sigma}}^{-1} \left(\widehat{\boldsymbol{\mu}}_{g's} - \widehat{\boldsymbol{\mu}}_{g''s}\right), \\
\widehat{\Delta}^3_{g'g''} &= \sum_{s=1}^{S-1} \frac{\widehat{\pi}_{g's} + \widehat{\pi}_{g''s}}{2} \left(\widehat{\boldsymbol{\tau}}_{g's} - \widehat{\boldsymbol{\tau}}_{g''s}\right)^\top \widehat{\mathbf{R}}^{-1} \left(\widehat{\boldsymbol{\tau}}_{g's} - \widehat{\boldsymbol{\tau}}_{g''s}\right),
\end{aligned}$$

with the unknown parameters simply replaced by their MLEs from Chapter 4. The asymptotic distribution of $\widehat{\Delta}_{g'g''}$ is derived in the following section.

5.3 Asymptotic Results

Consider the problem of constructing a statistical test of

$$H : \boldsymbol{\theta}_{g'} = \boldsymbol{\theta}_{g''} \qquad \text{against} \qquad K : \boldsymbol{\theta}_{g'} \neq \boldsymbol{\theta}_{g''}. \tag{5.2}$$

The following theorem derives a large-sample test of (5.2) using Theorem 4.4 in Chapter 4. Note that H is equivalent to $H' : \Delta_{g'g''} = 0$.

Theorem 5.2 *Suppose $\mathcal{P}^{(g')}$ and $\mathcal{P}^{(g'')}$ are mixed-variate populations defined by the general mixed-data models with respective parameters $\boldsymbol{\theta}_{g'}$ and $\boldsymbol{\theta}_{g''}$. Under $H : \boldsymbol{\theta}_{g'} = \boldsymbol{\theta}_{g''}$, then*

$$\frac{n_{g'.} n_{g''.}}{n_{g'.} + n_{g''.}} \widehat{\Delta}_{g'g''} \xrightarrow{\mathcal{L}} \chi^2_P, \tag{5.3}$$

when $\frac{n_{g'.}}{n_{g''.}} \to \delta$ as $n_{g'.} \to \infty$, $n_{g''.} \to \infty$, where $\delta < \infty$ and $P = P_1 + P_2 + P_3$ is the total number of unknown parameters as defined in Theorem 4.4.

Proof. The proof is very similar to that of Proposition 3.1 of Bar-Hen and Daudin (1995). Let $\boldsymbol{\theta}$ be the common value of $\boldsymbol{\theta}_{g'}$ and $\boldsymbol{\theta}_{g''}$ under H. Similar to Bar-Hen and Daudin (1995), a first-order Taylor series expansion of $\widehat{\Delta}_{g'g''}$ at a neighborhood of $(\boldsymbol{\theta}_{g'}, \boldsymbol{\theta}_{g''})$ yields

$$\begin{aligned}
\widehat{\Delta}_{g'g''} &= \Delta_{g'g''} + \sum_{g=g',g''} (\widehat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g)^\top \frac{\partial \Delta_{g'g''}}{\partial \boldsymbol{\theta}_g} + \frac{1}{2} \sum_{g=g',g''} (\widehat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g)^\top \frac{\partial^2 \Delta_{g'g''}}{\partial \boldsymbol{\theta}_g \partial \boldsymbol{\theta}_g^\top} (\widehat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g) \\
&\quad + (\widehat{\boldsymbol{\theta}}_{g'} - \boldsymbol{\theta}_{g'})^\top \frac{\partial^2 \Delta_{g'g''}}{\partial \boldsymbol{\theta}_{g'} \partial \boldsymbol{\theta}_{g''}^\top} (\widehat{\boldsymbol{\theta}}_{g''} - \boldsymbol{\theta}_{g''}) + \sum_{g=g',g''} o(\|\widehat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g\|) \\
&= (\widehat{\boldsymbol{\theta}}_{g'} - \widehat{\boldsymbol{\theta}}_{g''})^\top \mathcal{I}_P(\boldsymbol{\theta}) (\widehat{\boldsymbol{\theta}}_{g'} - \widehat{\boldsymbol{\theta}}_{g''}),
\end{aligned}$$

under H, where $o(\|\widehat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g\|) \xrightarrow{p} 0$ as $\widehat{\boldsymbol{\theta}}_g \to \boldsymbol{\theta}_g$ for $g = g', g''$, and $\mathcal{I}_P(\boldsymbol{\theta})$ is the expected Fisher information matrix based on all the observations. By

Theorem 4.4, it follows that

$$\sqrt{\frac{n_{g'.}n_{g''.}}{n_{g'.}+n_{g''.}}}(\widehat{\boldsymbol{\theta}}_{g'}-\boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}_P\left(\mathbf{0}, \frac{1}{1+\delta}\mathcal{I}_P^{-1}(\boldsymbol{\theta})\right),$$
$$\sqrt{\frac{n_{g'.}n_{g''.}}{n_{g'.}+n_{g''.}}}(\widehat{\boldsymbol{\theta}}_{g''}-\boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}_P\left(\mathbf{0}, \frac{\delta}{1+\delta}\mathcal{I}_P^{-1}(\boldsymbol{\theta})\right).$$

Hence,

$$\sqrt{\frac{n_{g'.}n_{g''.}}{n_{g'.}+n_{g''.}}}\mathcal{I}_P^{1/2}(\widehat{\boldsymbol{\theta}}_{g'}-\widehat{\boldsymbol{\theta}}_{g''}) \xrightarrow{\mathcal{L}} \mathcal{N}_P(\mathbf{0}, \mathbf{I}_P),$$

and the result follows immediately.

□

Corollary 5.2.1 *If $Q = 0$, then under the null hypothesis of no difference between $\mathcal{P}^{(g')}$ and $\mathcal{P}^{(g'')}$,*

$$\frac{n_{g'.}n_{g''.}}{n_{g'.}+n_{g''.}}\left(\widehat{\Delta}^1_{g'g''}+\widehat{\Delta}^2_{g'g''}\right) \xrightarrow{\mathcal{L}} \chi^2_{P_1+P_2}, \tag{5.4}$$

when $\frac{n_{g'.}}{n_{g''.}} \to \delta$ as $n_{g'.} \to \infty$, $n_{g''.} \to \infty$, where $\delta < \infty$.

Proof. If $Q = 0$, then the general mixed data model reduces to the general location model, so that this result is equivalent to Proposition 3.1 of Bar-Hen and Daudin (1995).

□

Corollary 5.2.2 *If $C = Q = 0$, then under the null hypothesis of no difference between $\mathcal{P}^{(g')}$ and $\mathcal{P}^{(g'')}$,*

$$\frac{n_{g'.}n_{g''.}}{n_{g'.}+n_{g''.}}\widehat{\Delta}^1_{g'g''} \xrightarrow{\mathcal{L}} \chi^2_{P_1}, \tag{5.5}$$

when $\frac{n_{g'.}}{n_{g''.}} \to \delta$ as $n_{g'.} \to \infty$, $n_{g''.} \to \infty$, where $\delta < \infty$.

Proof. This is equivalent to Corollary 3.1.1 of Bar-Hen and Daudin (1995).

□

Theorem 5.3 *If $S = 1$, then under the null hypothesis of no difference between $\mathcal{P}^{(g')}$ and $\mathcal{P}^{(g'')}$,*

$$\frac{n_{g'.}n_{g''.}}{n_{g'.}+n_{g''.}}\left(\widehat{\Delta}^2_{g'g''}+\widehat{\Delta}^3_{g'g''}\right) \xrightarrow{\mathcal{L}} \chi^2_{P_2+P_3}, \tag{5.6}$$

when $\frac{n_{g'.}}{n_{g''.}} \to \delta$ as $n_{g'.} \to \infty$, $n_{g''.} \to \infty$, where $\delta < \infty$.

Proof. The proof is similar to that of Theorem 5.2; hence, the details are omitted.

□

Corollary 5.3.1 *Under the null hypothesis of no difference between $\mathcal{P}^{(g')}$ and $\mathcal{P}^{(g'')}$, which are both defined by the grouped continuous model, it follows that*

$$\frac{n_{g'.}n_{g''.}}{n_{g'.}+n_{g''.}}\widehat{\Delta}^3_{g'g''} \xrightarrow{\mathcal{L}} \chi^2_{P_3}, \tag{5.7}$$

when $\frac{n_{g'.}}{n_{g''.}} \to \delta$ as $n_{g'.} \to \infty$, $n_{g''.} \to \infty$, where $\delta < \infty$.

Proof. The corollary is obtained directly from Theorem 5.3 by setting $C = 0$.

□

Remark 5.3.1 *Theorem 5.2 generalizes Proposition 3.1 of Bar-Hen and Daudin (1995) to the general mixed-data model. In fact, Proposition 3.1 is stated above as Corollary 5.2.1.*

Remark 5.3.2 *Theorem 5.3 is a two-sample test for mixed data distributed according to the conditional grouped continuous model. Similar tests based on likelihood ratio and generalized Wald statistics are discussed by Lapidus (1998, Chapter 4).*

Remark 5.3.3 *Corollary 5.3.1 is a two-sample test for the grouped continuous model.*

The level and power of the test described in Theorem 5.2 are evaluated through simulations in the next section.

5.4 Simulation Study

In the simulations, general mixed-data models with $C = L = Q = 1$ and $S = 2$ are considered. The parameter is then $\boldsymbol{\theta}_g^\top = (\pi_g, \boldsymbol{\mu}_g^\top, \sigma^2, \gamma, \beta, \tau_g)$, where $\boldsymbol{\mu}_g^\top = (\mu_{g1}, \mu_{g2})$ with μ_{gs} the sth state mean of Y_g, γ is the standardized cutpoint α for the latent variable Y_g^* underlying Z_g, and τ_g is the effect of state 1 on Z_g relative to that of state 2. Note that $\gamma = \alpha/\sqrt{1-\rho^2} - (\mu_2^*/\sqrt{1-\rho^2} - \beta\mu_2)$, $\beta = \rho/(\sigma\sqrt{1-\rho^2})$, and $\tau_g = \xi_g^*/\sqrt{1-\rho^2} - \beta\xi_g$, where $\xi_g = \mu_{g1} - \mu_2$, $\xi_g^* = \mu_{g1}^* - \mu_2^*$, with $\mu_2 = \mu_{g2}$, $\mu_2^* = \mu_{g2}^*$, and $\mu_{gs}^* = \mathrm{E}(Y_g^* \mid \mathbf{x}_g = \mathbf{x}_{(s)})$, for $g = 1, 2$, and $s = 1, 2$. Note also that $Z_g = 2$ if $Y_g^* > \alpha$ and $Z_g = 1$ if $Y_g^* \leq \alpha$. Similar to Bar-Hen and Daudin (1995), the following five cases are considered:

(0) no differences between populations with respect to all three variable types;

(a) there is difference between populations only with respect to nominal

vector $\mathbf{x}$;

(b) there is difference between populations only with respect to continuous variable Y;

(c) there is difference between populations only with respect to ordinal variable Z;

(d) populations are different with respect to all three variable types.

To assess the size and power of the χ^2 test in Theorem 5.2, random samples of sizes $(n_1, n_2) = (50, 25), (50, 100)$, and $(100, 100)$ were generated from the general mixed-data models with $(\sigma^2, \rho, \alpha)^\top = (1, .5, 1)^\top$ and location parameters $(p_g, \mu_{g1}, \mu_{g2}, \mu^*_{g1}, \mu^*_{g2})^\top$, $g = 1, 2$, given by (0) $(.5, 0, .5, 0, .5)^\top$ for both populations, (a) $(.5, 0, .5, 0, .5)^\top$ for population 1 and $(.75, 0, .5, 0, .5)^\top$ for population 2, (b) $(.5, 0, .5, 0, .5)^\top$ for population 1 and $(.5, .5, .5, 0, .5)^\top$ for population 2, (c) $(.5, 0, .5, 0, .5)^\top$ for population 1 and $(.5, 0, .5, .5, .5)^\top$ for population 2, and (d) $(.5, 0, .5, 0, .5)^\top$ for population 1 and $(.75, .5, .5, .5, .5)^\top$ for population 2.

Observe that case (0) is taken as giving the true parameter configurations for both populations under the null hypothesis $H : \Delta_{12} = 0$. For each combination of case and (n_1, n_2) above, 1000 replications were generated in S-PLUS. Hypothesis H is then rejected if and only if $n_1 n_2 \widehat{\Delta}_{12}/(n_1 + n_2) > \chi^2_{7,.05} = 14.1$, the 95th percentile of the χ^2 distribution with 7 degrees of freedom. Results of the simulated levels and powers of the test are displayed in Table 5.1.

Three observations are apparent from the table. First, the power of the test increases with the total sample size $n_1 + n_2$. Second, the test tends to be liberal when the total sample size is small, confirming an earlier finding reported by Bar-Hen and Daudin (1995). However, given large enough samples, the test is able to attain the nominal level. Finally, as was similarly reported by Bar-Hen and Daudin (1995), the power of the test is higher when differences exist with respect to all three variables than when the difference is only with respect to just one variable.

5.5 Example

In this section, the appendicitis data in Chapter 4 are revisited to illustrate the distance developed in previous sections. The data come from Koepsel *et al.* (1981) (also found on pp. 680-683 of Fisher and Van Bell, 1993) and concern the occurrence and non-occurrence of perforation of the appendix. Data from a total of 181 surgery patients are included in the analysis, and four variables are considered. The same data were analyzed by Nakanishi (1996) in the context of variable selection in mixed-data discriminant analysis.

For the purpose of this example, the same three variables studied in Chapter 4 are considered: the nominal vector $\mathbf{x}^\top = (X_1, X_2)$ correspond with the patient's perforation status, with $\mathbf{x} = \mathbf{x}_{(2)}$ if perforation is present and $\mathbf{x} = \mathbf{x}_{(1)}$ otherwise, the continuous variable Y representing the time in hours from physician contact to surgery, and the ordinal variable Z corresponding to the duration (long or short) of symptoms prior to physician contact. Patients

were grouped according to sex (i.e., male of female), and the interest is to see whether there is a difference between these two groups. As was done in Chapter 4, the waiting times to surgery were transformed using their natural logarithms. In addition, subjects with waiting times to surgery equal to 0 or exceeding 60 hours were not considered for the analysis.

The data set is summarized with respect to the discrete variables $\mathbf{x}$ and Z in Table 5.2. The values of the continuous variable Y are not shown in the table but can be obtained from Fisher and Van Bell (1993, p. 680). The general mixed-data model was fit to this data set and MLEs of the parameters were calculated using S-PLUS. These estimates are presented in Table 5.3.

From Table 5.3, $\widehat{\Delta}_{12}$ is found to be equal to 0.0396, and upon comparison with the 5% level critical value 14.1 obtained from the χ^2 distribution with 7 degrees of freedom, the test fails to reject the null hypothesis H that there is no difference due to sex. This conclusion agrees with Nakanishi's (1996) observations.

5.6 Discussion

In this chapter, a distance for mixed nominal, ordinal and continuous data was developed by applying the Kullback-Leibler divergence to the general mixed-data model. The distance so obtained can be considered as a generalization of the Mahalanobis distance to data with a mixture of nominal, ordinal and continuous variables. Moreover, it includes previous Mahalanobis-type distances developed by Bedrick *et al.* (2000) and Bar-Hen and Daudin (1995) as special

cases.

Asymptotic results regarding the maximum likelihood estimator of the distance were also discussed. Specifically, its asymptotic distribution when the populations are identical was derived, and large-sample tests were constructed based on it. The results of a simulation study on the level and power of the tests were reported as well. These results indicate that asymptotic tests were powerful enough to detect differences between populations and are able to maintain the nominal level given large enough samples. Finally, a real-data example was discussed to illustrate the method.

Table 5.1: *Empirical Size and Power of χ^2 Test in Theorem 5.2 for $C = L = Q = 1$ and $S = 2$ based on 1,000 Monte Carlo Samples.*

Source of Difference			Sample Size		
nominal	continuous	ordinal			
x	Y	Z	n_1	n_2	Power
	(0) $\Delta_{12} = \Delta_{21} = 0$				
No	No	No	50	25	0.112
No	No	No	50	100	0.109
No	No	No	100	100	0.054
	(a) $p_1 = 0.5, p_2 = 0.75$				
Yes	No	No	50	25	0.201
Yes	No	No	50	100	0.3
Yes	No	No	100	100	0.481
	(b) $\mu_{11} = 0, \mu_{21} = 0.5$				
No	Yes	No	50	25	0.146
No	Yes	No	50	100	0.193
No	Yes	No	100	100	0.275
	(c) $\mu^*_{11} = 0, \mu^*_{21} = 0.5$				
No	No	Yes	50	25	0.126
No	No	Yes	50	100	0.217
No	No	Yes	100	100	0.324
	(d) differences in all 3 variables				
Yes	Yes	Yes	50	25	0.287
Yes	Yes	Yes	50	100	0.483
Yes	Yes	Yes	100	100	0.733

NOTE: *The parameters under $H : \Delta_{12} = \Delta_{21} = 0$ (i.e., under case (0)) are $p_1 = p_2 = .5, \mu_{11} = \mu_{22} = \mu^*_{11} = \mu^*_{21} = 0, \mu_{12} = \mu_{22} = \mu^*_{12} = \mu^*_{22} = .5$ with $\sigma = 1, \rho = .5$. Note that* (a) *corresponds to the case where difference is only in* **x**, (b) *to difference in Y only,* (c) *to difference in Z only, and* (d) *to differences in all three.*

Table 5.2: *Three-Dimensional Array for the Appendicitis Data (Koepsel et al., 1981) classified by Sex.*

	Males		Females		
	Perforation		Perforation		
Duration	Yes	No	Yes	No	Total
> 24 hrs.	20	26	8	14	68
≤ 24 hrs.	5	61	5	42	113
Total	25	87	13	56	181

NOTE: *Shown are the numbers of surgery patients classified according to population (male or female), perforation state (s=1 if perforation is present and s=2 otherwise), and duration (Z=2 if duration exceeds 24 hrs. and Z=1 otherwise). The actual values of the time Y from diagnosis to surgery are found in Fisher and Van Bell (1993, p. 680).*

Table 5.3: *Maximum Likelihood Estimates of Parameters of General Mixed-Data Model for the Appendicitis Data classified by Sex.*

Parameter	Male Population	Female Population
$\widehat{p}$	0.2232 (0.039)	0.1884 (0.047)
$\widehat{\mu}_1$	1.2154 (0.211)	1.1908 (0.292)
$\widehat{\mu}_2$	1.5513 (0.113)	1.5972 (0.106)
$\widehat{\gamma}$	0.9022 (0.174)	1.0555 (0.276)
$\widehat{\beta}$	0.2448 (0.099)	0.2379 (0.144)
$\widehat{\tau}$	1.4365 (0.271)	1.0512 (0.237)
$\widehat{\sigma} = 1.0535\ (0.116)$		
$\widehat{\Delta}_{12} = \widehat{\Delta}_{21} = 0.0396$		

NOTE: *Shown are the maximum likelihood estimates of the general mixed-data model parameters for the male and female populations. The numbers in parentheses are the standard errors of the estimates. Also shown is the estimated generalized Mahalanobis distance between the two groups.*

Chapter 6

Concluding Remarks

6.1 Summary

The analysis of mixed data is not straightforward because of a lack of standard models for the joint distribution of the variables. Besides the ad-hoc approach of carrying out separate analyses for the discrete and continuous variables in the data, which are clearly deficient in many applications, a number of model-based alternatives have been previously proposed. These include the general location model for mixed nominal and continuous data and the conditional grouped continuous model for mixed data with ordinal and continuous variables. A number of issues concerning these models, in particular, and mixed data analyses, in general, were identified and addressed in the thesis.

This thesis focused on four main issues arising in mixed data analysis. The general approach taken in this thesis was a model-based one that relies on specifying a model for the joint distribution of the variables. Inferences are then developed for the parameters of the model. The approach is motivated by the need to account for the different measurement levels of the variables as well as the various associations among them. This approach is also preferable

to those that carry out separate analyses for discrete and continuous variables, in that it provides a systematic and non-ad hoc way of analyzing mixed data.

Chapter 2 tackled the problem of constructing global tests of location hypotheses in the context of mixed multinomial and continuous data. The likelihood ratio approach was used to derive tests in the one-sample and multi-sample settings after specifying a general location model for the joint distribution of the mixed-variable data. The approach allowed the problem to be treated from a multivariate perspective to simultaneously test both the discrete and continuous parameters of the model. One advantage of this approach is that it avoids the problem of multiple significance testing. Moreover, associations among the variables are accounted for, resulting in improved power performance of the tests (Pocock *et al.*, 1987). Unlike the tests previously proposed by Morales *et al.* (1998) which rely on asymptotic theory, the proposed likelihood ratio tests are all exact.

For the one-sample case, it was shown that the likelihood ratio test is both consistent and unbiased. A simulation study also showed that it performs quite competitively relative to the *separate test approach*, which carries out separate but simultaneous tests of the location parameters. The latter was shown to be not very powerful, especially in small samples.

The likelihood ratio tests in the multi-sample situation extend previous work by Afifi and Elashoff (1969) on the two-sample case. An attempt was made to relax the homogeneity assumption in the general location model, in the process extending the so-called Behrens-Fisher problem to the mixed data

case. Two situations where heterogeneity could arise in the general location model were considered. The same complications encountered in the case of continuous data were also identified in the mixed data case. It is an open question whether Behrens-Fisher solutions (Hussein and Carrière, 2001) can be adapted to the mixed data problem.

The proposed tests can be viewed as extensions to the mixed data setting of common likelihood ratio tests for continuous data in the one-sample case and of the classical multivariate analysis of variance problem in the multi-sample case (Seber, 1984; Mardia *et al.*, 1979).

In Chapter 3, an alternative estimation method was proposed for the grouped continuous model and its extension to mixed ordinal and continuous data, the so-called conditional grouped continuous model. The goal in this chapter is not to supplant the standard maximum likelihood approach but rather to devise an alternative practical method that strikes a balance between computational feasibility and statistical efficiency. The proposed method, based on the pairwise likelihood approach (Kuk and Nott, 2000), derives from Lindsay's (1988) concept of composite likelihood, which advocates simply pooling (or *compositing*) marginal (bivariate, in the case of pairwise likelihood approach) likelihoods to approximate the full likelihood. The composite likelihood function provides a single objective function, which is then maximized to obtain the estimates. Unlike maximum likelihood estimation, the maximum pairwise likelihood method is computationally simple, and unlike partition maximum likelihood methods (Bedrick *et al.*, 2000; Poon

and Lee, 1987), there is no need to combine multiple sets of estimates since it yields a single set of estimates. Like maximum likelihood estimators, maximum pairwise likelihood estimators were shown to be consistent and asymptotically normally distributed. This provides a route to constructing large-sample tests of hypotheses concerning the parameters of the model. Simulations showed that the estimates are quite accurate, yielding minimal bias and small root mean-squared errors.

A general model for mixed nominal, ordinal and continuous data called the *general mixed-data model*, was developed in Chapter 4. Despite the ubiquitousness of such data in practice, no model for the joint distribution of nominal, ordinal and continuous variables has yet been proposed. The new model is made up of two components: (1) a general location model for the joint distribution of the nominal and continuous variables, and (2) a conditional grouped continuous model for the joint distribution of the ordinal and continuous variables, given the nominal data. The *hybrid* model not only accounts for the ordinal information in the data but also incorporates associations between nominal and ordinal, nominal and continuous, and ordinal and continuous variables. It is flexible enough to be applicable to various types of mixed data and includes the general location and grouped continuous models as special cases. In this respect, the model provides a unified treatment of these two conventional mixed data models. Maximum likelihood and maximum pairwise likelihood methods were outlined for the model, and the asymptotic distributions of the corresponding estimators were derived. The

latter provides a more computationally feasible estimation method than the former, with only a minimal loss in efficiency. The asymptotic distributions of the maximum likelihood and maximum pairwise likelihood estimators were used to construct large-sample tests concerning the model parameters.

Finally, the general mixed-data model was used in Chapter 5 to develop a generalized Mahalanobis distance for mixed data. Because distance measures are common ingredients in many multivariate methods (e.g., cluster analysis, discrimination problems), developing one that can be used for mixed data has received a lot of attention in the literature (e.g., Krzanowski, 1993). However, none has yet been proposed for mixed data with ordinal, in addition to nominal and continuous, variables. The development of such a distance was the goal of Chapter 5. Following Bar-Hen and Daudin (1995), the Kullback-Leibler divergence was applied to the general mixed-data model to derive a distance measure for mixed nominal, ordinal and continuous data. The distance so obtained can be considered an extension of the Mahalanobis distance to mixed data. Moreover, previous generalizations of the Mahalanobis distance given by Bar-Hen and Daudin (1995) and Bedrick *et al.* (2000) were shown to be special cases of the new distance. Asymptotic distributions of the distance under the hypothesis of non-distinct groups were derived, and large-sample tests of hypotheses were constructed. Previous theoretical results due to Bar-Hen and Daudin (1995) were shown to be special cases of the new results. A simulation study was also undertaken to assess the performance of the tests in finite samples. The results of the simulations confirmed Bar-Hen and Daudin's

(1995) observations regarding the size and power of their tests.

6.2 Future Research

There are a number of important issues concerning mixed data analysis that can be pursued further from this thesis.

First, there is a need to relax the homogeneity assumption in the general location model. Although an attempt was made to incorporate heterogeneity for the multi-sample case in Chapter 2, no clear solution was found. Several approaches to incorporating heterogeneity in the state covariance matrices of the general location model have been previously proposed by Barnard *et al.* (2000) and Liu and Rubin (1998), among others. Such modifications to the general location model could be adopted in the construction of likelihood ratio tests for the multi-sample case as a possible solution to the Behrens-Fisher problem. With regard to the general mixed-data model, relaxing the homogeneity assumption will give more flexibility to the model in applications.

Second, the robustness of the grouped continuous model against the normality assumption for the latent variable distribution needs further study. Lee and Lam (1988) investigated this problem in the bivariate case, where they considered elliptical distributions for the latent variables, in general, and the bivariate t and *contaminated normal* distributions, in particular. A similar investigation was undertaken by Tan *et al.* (1999). Investigation in the general multivariate case needs to be pursued for a comprehensive study of the problem. As well, the family of elliptical distributions, which include the

multivariate t distribution, provides a way of handling heterogeneous state covariance structures in the general mixed-data model. Liu and Rubin's (1998) paper outlines a general approach of estimation for such model.

Third, although maximum pairwise likelihood estimates have been shown to have reasonably good performance in finite samples (Kuk and Nott, 2000; Heagerty and Lele, 1998), more extensive simulation studies still need to be undertaken to assess further their bias and efficiency for the grouped continuous model, especially in comparison with maximum likelihood and other competing estimates.

Fourth, it will be worthwhile to investigate how the general mixed-data model developed in this thesis can be applied in mixed data regression analysis. Regression analysis of mixed data has received rather sparse attention in the literature, with most of the developments having been done only recently (for example, Gueorguieva and Agresti, 2001; Geys *et al.*, 2001; Fitzmaurice and Laird, 1997; Catalano and Ryan, 1992). Further development in the more general mixed data case considered in this thesis will be a timely contribution to the literature.

Finally, there is a need to device theoretically sound methods of dealing with missing value problems in mixed data. Simply performing separate analyses on the quantitative and qualitative variables will not work because different sets of observations may be used in each analysis, and interpreting the results then becomes difficult. Little and Schluchter (1985), along with Belin *et al.* (1999), Fitzmaurice and Laird (1997), and Schafer (1997) addressed

this issue in the case of the general location model, where the missing data are assumed to be *missing at random* (MAR) (Little and Rubin, 1987). Lee and Chiu (1990) and Lee and Leung (1992) also studied the problem for the grouped continuous model. It may be possible to extend their methods to mixed nominal, ordinal, and continuous data. An approach that can deal with non-monotone and non-MAR missing data in this context should be considered.

Bibliography

Afifi, A. A. and R. M. Elashoff (1969). Multivariate two sample tests with dichotomous and continuous variables. I. The location model. *Annals of Mathematical Statistics* **40**, 290–298.

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley & Sons.

Agresti, A. (1990). *Categorical Data Analysis*. Wiley & Sons.

Anderberg, M. R. (1973). *Cluster Analysis for Applications*. Academic Press.

Anderson, J. A. and J. D. Pemberton (1985). The grouped continuous model for multivariate ordered categorical variables and covariate adjustment. *Biometrics* **41**, 875–885.

Anderson, J. A. and P. R. Philips (1981). Regression, discrimination and measurement models for ordered categorical variables. *Applied Statistics* **30**, 22–31.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. 2nd ed.. Wiley & Sons.

Bar-Hen, A. and J. J. Daudin (1995). Generalization of the Mahalanobis distance in the mixed case. *Journal of Multivariate Analysis* **53**, 332–342.

Bar-Hen, A. and J. J. Daudin (1998). Asymptotic distribution of Matusita's distance: Application to the location model. *Biometrika* **85**, 477–481.

Barnard, J., R. McCulloch and X-L. Meng (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10**, 1281–1311.

Bedrick, E. J., J. Lapidus and J. F. Powell (2000). Estimating the Mahalanobis distance from mixed continuous and discrete data. *Biometrics* **56**, 394–401.

Belin, T. R., M-Y. Hu, A. S. Young and O. Grusky (1999). Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. *Statistics in Medicine* **18**, 3123–3135.

Birdsall, M. A., C. M. Farquhar and H. D. White (1997). Association between polycystic ovaries and extent of artery disease in women having cardiac catheterization. *Annals of Internal Medicine* **126**, 32–35.

Bishop, Y. M. M., S. E. Fienberg and P. W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* **37**, 29–51.

Boos, D. D. (1992). On generalized score tests. *American Statistician* **46**(4), 327–333.

Catalano, P. J. (1997). Bivariate modeling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine* **16**, 883–900.

Catalano, P. J. and L. M. Ryan (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association* **87**, 651–658.

Cox, D. R. (1972). The analysis of multivariate binary data. *Applied Statistics* **21**, 113–120.

Cox, D. R. and N. Wermuth (1992). Response models for mixed binary and quantitative variables. *Biometrika* **79**, 441–461.

Cox, D. R. and N. Wermuth (1996). *Multivariate Dependencies: Models, Analysis and Interpretations.* Chapman & Hall.

Cox, N. R. (1974). Estimation of the correlation between a continuous and a discrete variable. *Biometrics* **30**, 171–178.

Crowder, M. J. (1986). On consistency and inconsistency of estimating equations. *Econometric Theory* **2**, 305–330.

Daudin, J. J. and A. Bar-Hen (1999). Selection in discriminant analysis with continuous and discrete variables. *Computational Statistics and Data Analysis* **32**, 161–175.

Davis, A. W. (1970). Exact distributions of Hotelling's generalized T_0^2. *Biometrika* **57**, 187–191.

de Leon, A. R. and K. C. Carrière (2000). On the one-sample location hypothesis for mixed bivariate data. *Communications in Statistics-Theory and Methods* **29**(11), 2573–2581.

de Leon, A. R. and K. C. Carrière (2001). Incorporating ordinal variables in the general location model: An extension. In: *2001 Proceedings of the Joint Statistical Meetings*. To appear.

Dillon, W. R. and M. Goldstein (1978). On the performance of some multinomial classification rules. *Journal of the American Statistical Association* **73**, 305–313.

Drasgow, F. (1986). Polychoric and polyserial correlations. In: *Encyclopedia of Statistical Sciences* (S. Kotz and N. L. Johnson, Eds.). pp. 68–74. Wiley & Sons.

Dyer, D. (1982). The convolution of generalized F distributions. *Journal of the American Statistical Association* **77**, 184–189.

Dykstra, R. L. (1970). Establishing the positive definiteness of the sample covariance matrix. *Annals of Mathematical Statistics* **41**, 2153–2154.

Edwards, D. (1995). *Introduction to Graphical Modelling*. Springer.

Faires, J. D. and R. L. Burden (1998). *Numerical Methods*. 2nd ed.. Brooks/Cole Pub. Co.

Fisher, L. D. and G. Van Bell (1993). *Biostatistics: A Methodolgy for the Health Sciences*. Wiley & Sons.

Fitzmaurice, G. M. and N. M. Laird (1997). Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics* **53**, 110–122.

Fletcher, R. and M. J. D. Powell (1964). A rapidly convergent method of minimization. *Computer Journal* **6**, 163–168.

Geisser, S. (1963). Multivariate analysis of variance for a special covariance case. *Journal of the American Statistical Association* **58**, 660–669.

Geys, H., M. M. Regan, P. J. Catalano and G. Molenberghs (2001). Two latent variable risk assessment approaches for mixed continuous and discrete outcomes from developmental toxicity data. *Journal of Agricultural, Biological, and Environmental Statistics* **6**, 340–355.

Gueorguieva, R. V. and A. Agresti (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association* **96**, 1102–1112.

Guyon, X. (1995). *Random Fields on a Network: Modeling, Statistics, and Applications*. Springer-Verlag.

Hannan, J. F. and R. F. Tate (1965). Estimation of the parameters for a multivariate normal distribution when one variable is dichotomized. *Biometrika* **52**, 664–668.

Harkness, W. L. (1965). Properties of the extended hypergeometric distribution. *Annals of Mathematical Statistics* **36**, 938–945.

Heagerty, P. J. and S. R. Lele (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* **93**, 1099–1111.

Hotelling, H. (1951). A generalized T-test and measure of multivariate dispersion. In: *Proceedings of Second Berkeley Symposium in Mathematical Statistics and Probability*. Vol. 1. pp. 23–41. University of California Press.

Hughes, D. T. and J. G. Saw (1972). Approximating percentage points of Hotelling's T_0^2 statistic. *Biometrika* **59**, 224–226.

Hussein, A. and K. C. Carrière (2001). Robustness of procedures for the Behrens-Fisher problems: Extension to bivariate normal mixtures. *Communications in Statistics-Simulation and Computation* **30**, 831–845.

Jóhannesson, B. and N. Giri (1995). On approximations involving the beta distribution. *Communications in Statistics-Simulation and Computation* **24**(2), 489–503.

Johnson, N. L. and S. Kotz (1970). *Continuous Univariate Distributions*. Vol. 2. Houghton Mifflin.

Johnson, N. L., S. Kotz and A. W. Kemp (1992). *Univariate Discrete Distributions*. 2nd ed.. Wiley & Sons.

Jordan, S. M. and K. Krishnamoorthy (1995). Confidence regions for the common mean vector of several multivariate normal populations. *Canadian Journal of Statistics* **23**, 283–297.

Koehler, K. J. and J. R. Wilson (1986). Chi-square tests for comparing vectors of proportions for several cluster samples. *Communications in Statistics-Theory and Methods* **15**, 2977–2990.

Koepsel, T. D., T. S. Inui and V. T. Farewell (1981). Factors affecting perforation in acute appendicitis. *Surgery, Gynecology and Obstetrics* **153**, 508–510.

Krishnaiah, P. R. and T. C. Chang (1972). On the exact distributions of the traces of $S_1(S_1 + S_2)^{-1}$ and $S_1S_2^{-1}$. *Sankhya-A* **34**, 153–160.

Krusińska, E. (1987). A valuation of state of object based on weighted Mahalanobis distance. *Pattern Recognition* **20**, 413–418.

Krusińska, E. and J. Liebhart (1988). The influence of outliers on discrmination of chronic obturative lung disease. *Methods of Information in Medicine* **27**, 167–176.

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics* **36**, 493–499.

Krzanowski, W. J. (1983). Distance between populations using mixed continuous and categorical variables. *Biometrika* **70**, 235–243.

Krzanowski, W. J. (1984). On the null distribution of distance between two groups, using mixed continuous and categorical variables. *Journal of Classification* **1**, 243–253.

Krzanowski, W. J. (1993). The location model for mixtures of categorical and continuous variables. *Journal of Classification* **10**, 25–49.

Kuk, A. Y. C. and D. J. Nott (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statistics and Probability Letters* **47**, 329–335.

Kullback, S. (1968). *Information Theory and Statistics*. 2nd ed.. Dover.

Kurczyński, T. W. (1970). Generalized distance and discrete variables. *Biometrics* **26**, 525–534.

Lapidus, J. (1998). Multivariate statistical methods using continuous and discrete data. PhD thesis. University of New Mexico.

Lauritzen, S. L. and N. Wermuth (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics* **17**, 31–54.

Lee, S-Y. (1985). Maximum likelihood stimation of polychoric correlations in $r \times s \times t$ contingency tables. *Journal of Statistical Computation and Simulation* **23**, 53–67.

Lee, S-Y. and K-M. Leung (1992). Estimation of multivariate polychoric and polyserial correlations with missing observations. *British Journal of Mathematical and Statistical Psychology* **45**, 225–238.

Lee, S-Y. and M-L. Lam (1988). Estimation of polychoric correlation with elliptical latent variables. *Journal of Statistical Computation and Simulation* **30**, 173–188.

Lee, S-Y. and S. K. Lau (1986). Estimation of polychoric correlation coefficient by the generalized least squares and associated methods. *Computational Statistics Quarterly* **3**, 61–73.

Lee, S-Y. and W-Y. Poon (1986). Maximum likelihood estimation of polyserial correlations. *Psychometrika* **51**, 113–121.

Lee, S-Y. and W-Y. Poon (1987). Two-step estimation of multivariate polychoric correlation. *Communications in Statistics-Theory and Methods* **16**(2), 307–320.

Lee, S-Y. and Y-M. Chiu (1990). Analysis of multivariate polychoric correlation models with incomplete data. *British Journal of Mathematical and Statistical Psychology* **43**, 145–154.

Lee, S-Y., W-Y. Poon and P. M. Bentler (1989). Simultaneous analysis of multivariate polytomous variates in several groups. *Psychometrika* **54**, 63–73.

Lehmann, E. L. (2000). *Elements of Large-Sample Theory*. Springer-Verlag.

Liang, K-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221–239.

Little, R. J. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. Wiley & Sons.

Little, R. J. and M. D. Schluchter (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72**, 496–512.

Liu, C. and D. B. Rubin (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika* **85**, 673–688.

Mardia, K. V. and P. J. Zemroch (1978). *Tables of the F- and Related Distributions with Algorithms*. Academic Press.

Mardia, K. V., J. T. Kent and J. M. Bibby (1979). *Multivariate Analysis*. Academic Press.

Martinson, E. O. and M. A. Hamdan (1971). Maximum likelihood and some other asymptotically efficient estimators of correlation in two-way contingency tables. *Journal of Statistical Computation and Simulation* **1**, 45–54.

Matusita, K. (1956). Decision rule, based on distance, for the classification problem. *Annals of the Institute of Statistical Mathematics* **16**, 305–315.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society-B* **42**, 109–142. (with discussion).

McDonald, R. P. and H. Swaminathan (1973). A simple matrix calculus with applications to multivariate analysis. *General Systems* **18**, 37–54.

McKeon, J. J. (1974). F approximation to the distribution of Hotelling's T_0^2. *Biometrika* **61**, 381–383.

Mijares, T. A. (1990). On the normal approximation to the Lawley-Hotelling trace criterion. *Biometrika* **77**, 443.

Montgomery, D. C. (1985). *Introduction to Statistical Quality Control.* Wiley & Sons.

Mood, A. M., F. A. Graybill and D. C. Boes (1974). *Introduction to the Theory of Statistics.* 3rd ed.. McGraw-Hill.

Morales, D., L. Pardo and K. Zografos (1998). Informational distances and related statistics in mixed continuous and categorical variables. *Journal of Statistical Planning and Inference* **75**, 47–63.

Morrison, D. F. (1971). The distribution of linear functions of independent F variates. *Journal of the American Statistical Association* **66**, 383–385.

Moustafa, M. D. (1957). Tests of hypotheses on a multivariate population, some of the variables being continuous and the rest categorical. *Institute of Statistics Mimeograph Series* 179. University of North Carolina-Chapel Hill.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology* **49**, 313–334.

Nakanishi, H. (1996). Distance between populations in a mixture of categorical and continuous variables. *Journal of the Japan Statistical Society* **26**, 221–230.

Nott, D. J. and T. Rydén (1999). Pairwise likelihood methods for inference in image models. *Biometrika* **86**, 661–676.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.

Ogawa, J., M. D. Moustafa and S. N. Roy (1957). On the asymptotic distribution of the likelihood ratio in some problems on mixed-variate populations. *Institute of Statistics Mimeograph Series* 180. University of North Carolina-Chapel Hill.

Olkin, I. and R. F. Tate (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics* **32**, 448–465. (correction in **36**, pp. 343-344).

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44**, 443–460.

Olsson, U., F. Drasgow and N. J. Dorans (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research* **14**, 485–500.

Olsson, U., F. Drasgow and N. J. Dorans (1982). The polyserial correlation coefficient. *Psychometrika* **47**, 337–347.

Parner, E. T. (2001). A composite likelihood approach to multivariate survival data. *Scandinavian Journal of Statistics* **28**, 295–302.

Pearson, E. S. and H. O. Hartley (1972). *Biometrika Tables for Statisticians.* Vol. 2. Cambridge University Press.

Pearson, K. (1904). Mathematical contribution to the theory of evolution. XIII. On the theory of contingency and its relation to association and normal correlation. *Biometrics Series* I. Drapers Co. Research Memoirs.

Pillai, K. C. S. and D. C. Young (1971). On the exact distribution of Hotelling's generalized T_0^2. *Journal of Multivariate Analysis* **1**, 90–107.

Pillai, K. C. S. and P. Sampson (1959). On Hotelling's generalization of T^2. *Biometrika* **46**, 160–168.

Plackett, R. L. (1954). A reduction formula for normal multivariate integrals. *Biometrika* **41**, 351–360.

Pocock, S. J., N. L. Geller and A. A. Tsiatis (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487–498.

Poon, W-Y. and S-Y. Lee (1987). Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients. *Psychometrika* **52**, 409–430. (correction in **53**, p. 301).

Poon, W-Y. and S-Y. Lee (1992). Statistical analysis of continuous and polytomous variables in several populations. *British Journal of Mathematical and Statistical Psychology* **45**, 139–149.

Poon, W-Y., S-Y. Lee, A. A. Afifi and P. M. Bentler (1990). Analysis of multivariate polytomous variates in several groups via the partition maximum likelihood approach. *Computational Statistics and Data Analysis* **10**, 17–27.

Raghunathan, T. E. and J. E. Grizzle (1995). A split questionnaire survey design. *Journal of the American Statistical Association* **90**, 55–63.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* 2nd ed.. Wiley & Sons.

Read, T. R. C. and N. A. C. Cressie (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer-Verlag.

Ronning, G. and M. Kukuk (1996). Efficient estimation of ordered probit models. *Journal of the American Statistical Association* **91**, 1120–1129.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489. (with discussion).

Sammel, M. D., L. M. Ryan and J. M. Legler (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society-B* **59**, 667–678.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall.

Seber, G. A. F. (1984). *Multivariate Observations.* Wiley & Sons.

Tallis, G. M. (1961). The moment generating function of the truncated multinormal distribution. *Journal of the Royal Statistical Society-B* **23**, 233–239.

Tallis, G. M. (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics* **18**, 342–353.

Tan, M., Y. Qu and J. S. Rao (1999). Robustness of the latent variable model for correlated binary data. *Biometrics* **55**, 258–263.

Tate, R. F. (1954). Correlation between a discrete and a continuous variable. *Annals of Mathematical Statistics* **25**, 603–607.

Tate, R. F. (1955). Applications of correlation models for biserial data. *Journal of the American Statistical Association* **50**, 1078–1095.

Truett, J., J. Cornfield and W. Kannel (1967). A multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Chronic Diseases* **20**, 511–524.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley & Sons.

Wiley, D. E. (1973). The identification problem for structural equation models with unmeasured variables. In: *Structural Equation Models in the Social Sciences* (A. S. Goldberger and O. D. Duncan, Eds.). pp. 69–83. Seminar Press.

Appendix A

Review of Optimization Methods

Numerical optimization methods used in the thesis for maximizing the likelihood or log-likelihood functions are briefly reviewed in this appendix.

Consider maximizing a log-likelihood or pseudo log-likelihood function $\ell(\boldsymbol{\theta}) \equiv \ell$ with respect to $\boldsymbol{\theta}$. The general form of the updating formula is given by

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \widehat{\boldsymbol{\theta}}^{(t)} + \alpha^{(t)}\mathbf{d}^{(t)},$$

for $t = 0, 1, \cdots$, where $\widehat{\boldsymbol{\theta}}^{(t)}$ is the estimate of the parameter $\boldsymbol{\theta}_{P\times 1}$ at iteration t, $\alpha^{(t)}$ is the *step size* at iteration t, and $\mathbf{d}^{(t)}$ is the $P \times 1$ *directional vector*.

A.1 Determining the Direction

The estimate $\widehat{\boldsymbol{\theta}}^{(t+1)}$ is said to be *acceptable* if $\widehat{\ell}^{(t+1)} > \widehat{\ell}^{(t)}$, with $\widehat{\ell}^{(t)}$ the function ℓ evaluated at $\widehat{\boldsymbol{\theta}}^{(t)}$. The vector $\mathbf{d}^{(t)}$ is generally chosen to be $\mathbf{d}^{(t)} = -\mathbf{G}^{-1}\mathbf{s}(\widehat{\boldsymbol{\theta}}^{(t)})$, where $\mathbf{G}$ is a negative definite matrix and $\mathbf{s}(\widehat{\boldsymbol{\theta}}^{(t)})$ is the *score vector* $\mathbf{s}(\boldsymbol{\theta}) = \partial\ell/\partial\boldsymbol{\theta}$, evaluated at $\widehat{\boldsymbol{\theta}}^{(t)}$. For minimization problems, $\mathbf{G}$ must be positive

definite.

A.2 Choosing a Step Size

There are three ways to choose the step size $\alpha^{(t)}$. The easiest is to fix it at some value α for all t. Another is to undertake a *linear search* as follows:

(i) find an acceptable $\mathbf{d}^{(t)}$;

(ii) find α^* such that $\alpha^* = \text{argmax}_\alpha \ell(\widehat{\boldsymbol{\theta}}^{(t)} + \alpha \mathbf{d}^{(t)})$, by, say, approximating $\ell(\widehat{\boldsymbol{\theta}}^{(t)} + \alpha \mathbf{d}^{(t)})$ with a polynomial in α.

Still another method is the so-called *step halving*. If $\ell(\widehat{\boldsymbol{\theta}}^{(t+1)}) < \ell(\widehat{\boldsymbol{\theta}}^{(t)})$, the method halves $\alpha^{(t)}$ until $\ell(\widehat{\boldsymbol{\theta}}^{(t+1)}) > \ell(\widehat{\boldsymbol{\theta}}^{(t)})$.

A.3 Survey of Basic Methods

The *steepest ascent method* takes $\mathbf{G} = -\mathbf{I}$, $\mathbf{d}^{(t)} = \mathbf{s}(\widehat{\boldsymbol{\theta}}^{(t)})$, so that the updating formula becomes

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \widehat{\boldsymbol{\theta}}^{(t)} + \alpha^{(t)} \mathbf{s}(\widehat{\boldsymbol{\theta}}^{(t)}).$$

This method approaches the true $\widehat{\boldsymbol{\theta}} = \lim_{t \to \infty} \widehat{\boldsymbol{\theta}}^{(t)}$ quite quickly but slows down as it gets closer to $\widehat{\boldsymbol{\theta}}$.

A better alternative is the *Newton-Raphson method*, which uses the following updating formula:

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \widehat{\boldsymbol{\theta}}^{(t)} - \mathbf{H}^{-1}(\widehat{\boldsymbol{\theta}}^{(t)}) \mathbf{s}(\widehat{\boldsymbol{\theta}}^{(t)}),$$

where $\mathbf{H}(\widehat{\boldsymbol{\theta}}^{(t)})$ is the *Hessian matrix* evaluated at $\widehat{\boldsymbol{\theta}}^{(t)}$, also referred to as the *observed information matrix*. The main drawback of this method is that it needs very good initial estimates to converge.

Modifications to the Newton-Raphson method have been proposed in cases where the Hessian matrix is not available or is not negative definite. These variants of the method include the *Levenberg-Marquardt* and *quasi-Newton* methods. The latter method falls under the general heading of *variable metric* or *secant methods*.

Another closely related method is the *Fisher scoring method*, which uses the *expected information matrix* $\mathcal{I}(\boldsymbol{\theta}) = \mathrm{E}[-\mathbf{H}(\boldsymbol{\theta})]$, evaluated at $\widehat{\boldsymbol{\theta}}^{(t)}$, in place of $-\mathbf{H}(\widehat{\boldsymbol{\theta}}^{(t)})$.

Finally, a method that estimates the Hessian matrix adaptively is given by the *Fletcher-Powell algorithm* (also known as *Davidson-Fletcher-Powell method*). It estimates the Hessian matrix $\widehat{\mathbf{H}}^{(t)} \equiv \mathbf{H}(\widehat{\boldsymbol{\theta}}^{(t)})$ as follows:

$$\widehat{\mathbf{H}}^{(t+1)} = \widehat{\mathbf{H}}^{(t)} + \frac{\mathbf{a}^{(t)}(\mathbf{a}^{(t)})^\top}{(\mathbf{a}^{(t)})^\top \mathbf{a}^{(t)}} - \frac{\widehat{\mathbf{H}}^{(t)}\mathbf{b}^{(t)}(\mathbf{b}^{(t)})^\top \widehat{\mathbf{H}}^{(t)}}{(\mathbf{b}^{(t)})^\top \widehat{\mathbf{H}}^{(t)} \mathbf{b}^{(t)}},$$

where $\mathbf{a}^{(t)} = \widehat{\boldsymbol{\theta}}^{(t)} - \widehat{\boldsymbol{\theta}}^{(t-1)}$ and $\mathbf{b}^{(t)} = \mathbf{s}(\widehat{\boldsymbol{\theta}}^{(t)}) - \mathbf{s}(\widehat{\boldsymbol{\theta}}^{(t-1)})$. The initial estimate is usually taken to be $\widehat{\mathbf{H}}^{(0)} = \mathbf{I}$.

For optimization problems, crude estimates of $\mathbf{H}(\boldsymbol{\theta})$ (e.g., $\mathbf{I}$) often suffice, but may result in extremely poor estimates of the covariance matrix of $\widehat{\boldsymbol{\theta}}$. More details regarding these methods are found in Faires and Burden (1998).

Appendix B

S-PLUS Programs Used in the Thesis

B.1 Calculation of Critical Values in Table 2.1

```
pnull←.3
ssize←15
alpha←0.01
pvalue←function(n,p,c)
{
   prob←0
   for (i in 1:(n-1))
   {
      A←((((i/n)^i)*((1-(i/n))^(n-i)))/((p^i)*((1-p)^(n-i))))^(2/n)
      const←1-p^n-(1-p)^n
      prob←prob+(1-pf(((n-2)*(c/A-1)/2),2,(n-2)))*dbinom(i,n,p)/const
   }
   pvalue←prob
   pvalue
}
critical1←qf(0.95,2,(ssize-2))
critical2←qf(0.5,2,(ssize-2))
margin←0.5
while(margin>0.000000001)
{
      critical←(critical1+critical2)/2
      ptail←pvalue(ssize,pnull,critical)
      error←ptail-alpha
      margin←abs(error)
      if (error<0)
```

```
        {
        critical1←critical
        }
        else
        {
        critical2←critical
        }
}
critical
```

B.2 Calculation of Power Values in Tables 2.2 and 2.3

```
pnull←.3
mean1null←50
mean2null←25
sigma←5
sigma.sq←sigma^2
size.vec←c(15,25)
alpha←0.01
alpha1←alpha/3
pvalue←function(n,p,c)
{
   prob←0
   for (i in 1:(n-1))
   {
      A←((((i/n)^i)*((1-(i/n))^(n-i)))/((p^i)*((1-p)^(n-i))))^(2/n)
      const←1-p^n-(1-p)^n
      prob←prob+(1-pf(((n-2)*(c/A-1)/2),2,(n-2)))*dbinom(i,n,p)/const
   }
   pvalue←prob
   pvalue
}
pvalue1←function(n,p,pr,c,scale,mu1,mu2,mu1null,mu2null)
{
   prob←0
   for (i in 1:(n-1))
   {
      A←((((i/n)^i)*((1-(i/n))^(n-i)))/((pr^i)*((1-pr)^(n-i))))^(2/n)
      const←1-p^n-(1-p)^n
      nc←(i*(mu1-mu1null)^2+(n-i)*(mu2-mu2null)^2)/scale
```

```
      prob←prob+(1-pf(((n-2)*(c/A-1)/2),2,(n-2),nc))*dbinom(i,n,p)/const
   }
   pvalue1←prob
   pvalue1
}
par1←c(.3,50,25)
par2←c(.35,50,25)
par3←c(.35,52.5,22.5)
par4←c(.4,55,22.5)
par5←c(.4,55,20)
param←rbind(par1,par2,par3,par4,par5)
power1.lrt←matrix(0,5,2)
power1.mt←matrix(0,5,2)
for (i in 1:5)
{
   ptrue←param[i,1]
   truemean1←param[i,2]
   truemean2←param[i,3]
   for (j in 1:2)
   {
      ssize←size.vec[j]
      critical1←qf(0.95,2,(ssize-2))
      critical2←qf(0.5,2,(ssize-2))
      margin←0.5
      while(margin>0.000000001)
      {
         critical←(critical1+critical2)/2
         ptail←pvalue(ssize,pnull,critical)
         error←ptail-alpha
         margin←abs(error)
         if (error<0)
         {
           critical1←critical
         }
         else
         {
           critical2←critical
         }
      }
      power1.lrt[i,j]←power1.lrt[i,j]+pvalue1(ssize,ptrue,pnull,critical,
      sigma.sq,truemean1,truemean2,mean1null,mean2null)
      count1←0
```

```
count2←0
set.seed(493)
N←10000
for (k in 1:N)
{
   n1←rbinom(1,ssize,ptrue)
   while(n1<1 || n1>(ssize-1))
   {
     n1←rbinom(1,ssize,ptrue)
   }
   n2←ssize-n1
   x1←rnorm(n1,mean=truemean1,sd=sigma)
   x2←rnorm(n2,mean=truemean2,sd=sigma)
   xbar1←mean(x1)
   xbar2←mean(x2)
   if (n1==1)
   {
     varx1←0
     varx2←var(x2)
   }
   else if (n1==(ssize-1))
   {
     varx1←var(x1)
     varx2←0
   }
   else
   {
     varx1←var(x1)
     varx2←var(x2)
   }
   pooled.var←((n1-1)*varx1+(n2-1)*varx2)/(ssize-2)
   test3←binom.test(n1,ssize,p=pnull,alternative="two.sided")
   stat.test1←(xbar1-mean1null)*sqrt(n1)/sqrt(pooled.var)
   stat.test2←(xbar2-mean2null)*sqrt(n2)/sqrt(pooled.var)
   if (abs(stat.test1)>qt((1-(alpha1/2)),(ssize-2))
   || abs(stat.test2)>qt((1-(alpha1/2)),(ssize-2))
   || test3$p.value<(alpha1/2))
   {
     count1←count1+1
   }
   else
   {
```

```
            count1←count1
          }
        }
        power1.mt[i,j]←power1.mt[i,j]+count1/N
    }
}
power1.mat←cbind(power1.lrt,power1.mt)
```

B.3 Calculation of Bias and RMSE in Tables 3.1 to 3.4

```
dim←3
ssize←50
cut1.1←.5
cut2.1←-.75
cut2.2←.1
cut3.1←-.25
cut3.2←.3
cut3.3←1
SIG←matrix(data=c(1,.8,.3,.8,1,.4,.3,.4,1),byrow=T,nrow=3,ncol=3)
reps←50
mplest.vec←matrix(0,nrow=reps,ncol=9)
for (j in 1:reps)
{
   set.seed(j*9)
   latent.dat←matrix(rnorm(ssize*dim),ncol=dim)%*%chol(SIG1)
   z.mat←matrix(0,nrow=ssize,ncol=dim)
   for (i in 1:ssize)
   {
      if (latent.dat[i,1]<=cut1.1)
      {
         z.mat[i,1]←1
      }
      else if (latent.dat[i,1]>cut1.1)
      {
         z.mat[i,1]←2
      }
      if (latent.dat[i,2]<=cut2.1)
      {
         z.mat[i,2]←1
      }
```

```
   else if (latent.dat[i,2]<=cut2.2)
   {
      z.mat[i,2]←2
   }
   else if (latent.dat[i,2]>cut2.2)
   {
      z.mat[i,2]←3
   }
   if (latent.dat[i,3]<=cut3.1)
   {
      z.mat[i,3]←1
   }
   else if (latent.dat[i,3]<=cut3.2)
   {
      z.mat[i,3]←2
   }
   else if (latent.dat[i,3]<=cut3.3)
   {
      z.mat[i,3]←3
   }
   else if (latent.dat[i,3]>cut3.3)
   {
      z.mat[i,3]←4
   }
}
n1.1.1←0
n1.1.2←0
n1.1.3←0
n1.1.4←0
n1.2.1←0
n1.2.2←0
n1.2.3←0
n1.2.4←0
n1.3.1←0
n1.3.2←0
n1.3.3←0
n1.3.4←0
n2.1.1←0
n2.1.2←0
n2.1.3←0
n2.1.4←0
n2.2.1←0
```

```
n2.2.2←0
n2.2.3←0
n2.2.4←0
n2.3.1←0
n2.3.2←0
n2.3.3←0
n2.3.4←0
for (i in 1:ssize)
{
   if (z.mat[i,1]==1)
   {
      if (z.mat[i,2]==1)
      {
        if (z.mat[i,3]==1)
        {
          n1.1.1←n1.1.1+1
        }
        else if (z.mat[i,3]==2)
        {
          n1.1.2←n1.1.2+1
        }
        else if (z.mat[i,3]==3)
        {
          n1.1.3←n1.1.3+1
        }
        else if (z.mat[i,3]==4)
        {
          n1.1.4←n1.1.4+1
        }
      }
      else if (z.mat[i,2]==2)
      {
        if (z.mat[i,3]==1)
        {
          n1.2.1←n1.2.1+1
        }
        else if (z.mat[i,3]==2)
        {
          n1.2.2←n1.2.2+1
        }
        else if (z.mat[i,3]==3)
        {
```

```
        n1.2.3←n1.2.3+1
      }
      else if (z.mat[i,3]==4)
      {
        n1.2.4←n1.2.4+1
      }
    }
    else if (z.mat[i,2]==3)
    {
      if (z.mat[i,3]==1)
      {
        n1.3.1←n1.3.1+1
      }
      else if (z.mat[i,3]==2)
      {
        n1.3.2←n1.3.2+1
      }
      else if (z.mat[i,3]==3)
      {
        n1.3.3←n1.3.3+1
      }
      else if (z.mat[i,3]==4)
      {
        n1.3.4←n1.3.4+1
      }
    }
  }
  else if (z.mat[i,1]==2)
  {
    if (z.mat[i,2]==1)
    {
      if (z.mat[i,3]==1)
      {
        n2.1.1←n2.1.1+1
      }
      else if (z.mat[i,3]==2)
      {
        n2.1.2←n2.1.2+1
      }
      else if (z.mat[i,3]==3)
      {
        n2.1.3←n2.1.3+1
```

```
  }
  else if (z.mat[i,3]==4)
  {
    n2.1.4←n2.1.4+1
  }
}
else if (z.mat[i,2]==2)
{
  if (z.mat[i,3]==1)
  {
    n2.2.1←n2.2.1+1
  }
  else if (z.mat[i,3]==2)
  {
    n2.2.2←n2.2.2+1
  }
  else if (z.mat[i,3]==3)
  {
    n2.2.3←n2.2.3+1
  }
  else if (z.mat[i,3]==4)
  {
    n2.2.4←n2.2.4+1
  }
}
else if (z.mat[i,2]==3)
{
  if (z.mat[i,3]==1)
  {
    n2.3.1←n2.3.1+1
  }
  else if (z.mat[i,3]==2)
  {
    n2.3.2←n2.3.2+1
  }
  else if (z.mat[i,3]==3)
  {
    n2.3.3←n2.3.3+1
  }
  else if (z.mat[i,3]==4)
  {
    n2.3.4←n2.3.4+1
```

```
        }
      }
    }
}
minuspairwiseloglike<-function(par)
{
   if (par[1]<-1 || par[1]>1)
   {
      par[1]<-init[1]
   }
   else
   {
      par[1]<-par[1]
   }
   if (par[2]<-1 || par[2]>1)
   {
      par[2]<-init[2]
   }
   else
   {
      par[2]<-par[2]
   }
   if (par[3]<-1 || par[3]>1)
   {
      par[3]<-init[3]
   }
   else
   {
      par[3]<-par[3]
   }
   term.a<-(n1.1.1+n1.1.2+n1.1.3+n1.1.4)
   *log(pmvnorm(c(par[4],par[5]),rho=par[1]))
   term.b<-(n1.2.1+n1.2.2+n1.2.3+n1.2.4)
   *log(pmvnorm(c(par[4],par[6]),rho=par[1])
   -pmvnorm(c(par[4],par[5]),rho=par[1]))
   term.c<-(n1.3.1+n1.3.2+n1.3.3+n1.3.4)*log(pnorm(par[4])
   -pmvnorm(c(par[4],par[6]),rho=par[1]))
   term.d<-(n2.1.1+n2.1.2+n2.1.3+n2.1.4)*log(pnorm(par[5])
   -pmvnorm(c(par[4],par[5]),rho=par[1]))
   term.e<-(n2.2.1+n2.2.2+n2.2.3+n2.2.4)*log(pnorm(par[6])
   -pnorm(par[5])-pmvnorm(c(par[4],par[6]),rho=par[1])
   +pmvnorm(c(par[4],par[5]),rho=par[1]))
```

```
term.f←(n2.3.1+n2.3.2+n2.3.3+n2.3.4)*log(1-pnorm(par[6])
-pnorm(par[4])+pmvnorm(c(par[4],par[6]),rho=par[1]))
term.g←(n1.1.1+n1.2.1+n1.3.1)
*log(pmvnorm(c(par[4],par[7]),rho=par[2]))
term.h←(n1.1.2+n1.2.2+n1.3.2)
*log(pmvnorm(c(par[4],par[8]),rho=par[2])
-pmvnorm(c(par[4],par[7]),rho=par[2]))
term.i←(n1.1.3+n1.2.3+n1.3.3)
*log(pmvnorm(c(par[4],par[9]),rho=par[2])
-pmvnorm(c(par[4],par[8]),rho=par[2]))
term.j←(n1.1.4+n1.2.4+n1.3.4)*log(pnorm(par[4])
-pmvnorm(c(par[4],par[9]),rho=par[2]))
term.k←(n2.1.1+n2.2.1+n2.3.1)*log(pnorm(par[7])
-pmvnorm(c(par[4],par[7]),rho=par[2]))
term.l←(n2.1.2+n2.2.2+n2.3.2)*log(pnorm(par[8])
-pnorm(par[7])-pmvnorm(c(par[4],par[8]),rho=par[2])
+pmvnorm(c(par[4],par[7]),rho=par[2]))
term.m←(n2.1.3+n2.2.3+n2.3.3)*log(pnorm(par[9])
-pnorm(par[8])-pmvnorm(c(par[4],par[9]),rho=par[2])
+pmvnorm(c(par[4],par[8]),rho=par[2]))
term.n←(n2.1.4+n2.2.4+n2.3.4)*log(1-pnorm(par[9])
-pnorm(par[4])+pmvnorm(c(par[4],par[9]),rho=par[2]))
term.o←(n1.1.1+n2.1.1)*log(pmvnorm(c(par[5],par[7]),rho=par[3]))
term.p←(n1.1.2+n2.1.2)*log(pmvnorm(c(par[5],par[8]),rho=par[3])
-pmvnorm(c(par[5],par[7]),rho=par[3]))
term.q←(n1.1.3+n2.1.3)*log(pmvnorm(c(par[5],par[9]),rho=par[3])
-pmvnorm(c(par[5],par[8]),rho=par[3]))
term.r←(n1.1.4+n2.1.4)*log(pnorm(par[5])
-pmvnorm(c(par[5],par[9]),rho=par[3]))
term.s←(n1.2.1+n2.2.1)*log(pmvnorm(c(par[6],par[7]),rho=par[3])
-pmvnorm(c(par[5],par[7]),rho=par[3]))
term.t←(n1.2.2+n2.2.2)*log(pmvnorm(c(par[6],par[8]),rho=par[3])
-pmvnorm(c(par[6],par[7]),rho=par[3])
-pmvnorm(c(par[5],par[8]),rho=par[3])
+pmvnorm(c(par[5],par[7]),rho=par[3]))
term.u←(n1.2.3+n2.2.3)*log(pmvnorm(c(par[6],par[9]),rho=par[3])
-pmvnorm(c(par[6],par[8]),rho=par[3])
-pmvnorm(c(par[5],par[9]),rho=par[3])
+pmvnorm(c(par[5],par[8]),rho=par[3]))
term.v←(n1.2.4+n2.2.4)*log(pnorm(par[6])
-pmvnorm(c(par[6],par[9]),rho=par[3])
-pnorm(par[5])+pmvnorm(c(par[5],par[9]),rho=par[3]))
```

```
      term.w<-(n1.3.1+n2.3.1)*log(pnorm(par[7])
      -pmvnorm(c(par[6],par[7]),rho=par[3]))
      term.x<-(n1.3.2+n2.3.2)*log(pnorm(par[8])
      -pnorm(par[7])-pmvnorm(c(par[6],par[8]),rho=par[3])
      +pmvnorm(c(par[6],par[7]),rho=par[3]))
      term.y<-(n1.3.3+n2.3.3)*log(pnorm(par[9])
      -pnorm(par[8])-pmvnorm(c(par[6],par[9]),rho=par[3])
      +pmvnorm(c(par[6],par[8]),rho=par[3]))
      term.z<-(n1.3.4+n2.3.4)*log(1-pnorm(par[9])
      -pnorm(par[6])+pmvnorm(c(par[6],par[9]),rho=par[3]))
      minuspairwiseloglike<--(term.a+term.b+term.c+term.d+term.e
      +term.f+term.g+term.h+term.i+term.j+term.k+term.l+term.m
      +term.n+term.o+term.p+term.q+term.r+term.s+term.t+term.u
      +term.v+term.w+term.x+term.y+term.z)
   }
   gcm.mplest<-nlmin(f=minuspairwiseloglike,
   x=init,max.iter=100,max.fcal=100)
   mplest.vec[j,]<-gcm.mplest$x
}
true.pars<-c(SIG[1,2],SIG[1,3],SIG[2,3],cut1.1,
cut2.1,cut2.2,cut3.1,cut3.2,cut3.3)
ave.mplest<-c(mean(mplest.vec[,1]),mean(mplest.vec[,2]),
mean(mplest.vec[,3]),mean(mplest.vec[,4]),mean(mplest.vec[,5]),
mean(mplest.vec[,6]),mean(mplest.vec[,7]),
mean(mplest.vec[,8]),mean(mplest.vec[,9]))
rmse.vec<-sqrt(c(var(mplest.vec[,1],unbiased=F),
var(mplest.vec[,2],unbiased=F),var(mplest.vec[,3],unbiased=F),
var(mplest.vec[,4],unbiased=F),var(mplest.vec[,5],unbiased=F),
var(mplest.vec[,6],unbiased=F),var(mplest.vec[,7],unbiased=F),
var(mplest.vec[,8],unbiased=F),var(mplest.vec[,9],unbiased=F)))
bias.mplest<-true.pars-ave.mplest
rel.bias<-(bias.mplest/true.pars)*100
```

B.4 Calculation of Power Values in Table 5.1

```
ssize1<-50
ssize2<-100
p1<-.5
p2<-.75
mu1.1<-0
mu2.1<-.5
mu.2<-.5
```

```
mu1.1.star←0
mu2.1.star←.5
mu.2.star←.5
xi1.star←mu1.1.star-mu.2.star
xi2.star←mu2.1.star-mu.2.star
xi1←mu1.1-mu.2
xi2←mu2.1-mu.2
sigma←1
rho←.5
alpha←1
beta←rho/(sigma*sqrt(1-rho^2))
gamma←alpha/sqrt(1-rho^2)-mu.2.star/sqrt(1-rho^2)+beta*mu.2
tau1←xi1.star/sqrt(1-rho^2)-beta*xi1
tau2←xi2.star/sqrt(1-rho^2)-beta*xi2
level←.05
count←0
rep←1000
for (i in 1:rep)
{
   n1.1←rbinom(1,ssize1,p1)
   while(n1.1<1 || n1.1>(ssize1-1))
   {
      n1.1←rbinom(1,ssize1,p1)
   }
   n1.2←ssize1-n1.1
   n2.1←rbinom(1,ssize2,p2)
   while(n2.1<1 || n2.1>(ssize2-1))
   {
      n2.1←rbinom(1,ssize2,p2)
   }
   n2.2←ssize2-n2.1
   p1.hat←n1.1/ssize1
   q1.hat←1-p1.hat
   p2.hat←n2.1/ssize2
   q2.hat←1-p2.hat
   y1.1←rnorm(n1.1,mean=mu1.1,sd=sigma)
   y1.2←rnorm(n1.2,mean=mu1.2,sd=sigma)
   y2.1←rnorm(n2.1,mean=mu2.1,sd=sigma)
   y2.2←rnorm(n2.2,mean=mu2.2,sd=sigma)
   mu1.1.hat←mean(y1.1)
   mu1.2.hat←mean(y1.2)
   mu2.1.hat←mean(y2.1)
```

```
mu2.2.hat←mean(y2.2)
sigmasq.hat←((n1.1-1)*var(y1.1)+(n1.2-1)*var(y1.2)
+(n2.1-1)*var(y2.1)+(n2.2-1)*var(y2.2))/(ssize1+ssize2)
y1.1.star←(rnorm(n1.1,mean=0,sd=1)+mu.2.star/sqrt(1-rho^2)
-beta*mu.2+tau1+beta*y1.1)*sqrt(1-rho^2)
y1.2.star←(rnorm(n1.2,mean=0,sd=1)+mu.2.star/sqrt(1-rho^2)
-beta*mu.2+beta*y1.2)*sqrt(1-rho^2)
y2.1.star←(rnorm(n2.1,mean=0,sd=1)+mu.2.star/sqrt(1-rho^2)
-beta*mu.2+tau2+beta*y2.1)*sqrt(1-rho^2)
y2.2.star←(rnorm(n2.2,mean=0,sd=1)+mu.2.star/sqrt(1-rho^2)
-beta*mu.2+beta*y2.2)*sqrt(1-rho^2)
z1.1←c(rep(0,n1.1))
z1.2←c(rep(0,n1.2))
for (i in 1:n1.1)
{
   if (y1.1.star[i]>alpha)
   {
      z1.1[i]←2
   }
   else
   {
      z1.1[i]←1
   }
}
for (i in 1:n1.2)
{
   if (y1.2.star[i]>alpha)
   {
      z1.2[i]←2
   }
   else
   {
      z1.2[i]←1
   }
}
z2.1←c(rep(0,n2.1))
z2.2←c(rep(0,n2.2))
for (i in 1:n2.1)
{
   if (y2.1.star[i]>alpha)
   {
      z2.1[i]←2
```

```
    }
    else
    {
       z2.1[i]←1
    }
}
for (i in 1:n2.2)
{
    if (y2.2.star[i]>alpha)
    {
       z2.2[i]←2
    }
    else
    {
       z2.2[i]←1
    }
}
count1.1.1←0
count1.1.2←0
noty1.1.1←c(rep(0,n1.1))
noty1.1.2←c(rep(0,n1.1))
for (i in 1:n1.1)
{
    if (z1.1[i]==1)
    {
       count1.1.1←count1.1.1+1
       noty1.1.1[count1.1.1]←y1.1[i]
    }
    else
    {
       count1.1.2←count1.1.2+1
       noty1.1.2[count1.1.2]←y1.1[i]
    }
}
count1.2.1←0
count1.2.2←0
noty1.2.1←c(rep(0,n1.2))
noty1.2.2←c(rep(0,n1.2))
for (i in 1:n1.2)
{
    if (z1.2[i]==1)
    {
```

```
      count1.2.1←count1.2.1+1
      noty1.2.1[count1.2.1]←y1.2[i]
   }
   else
   {
      count1.2.2←count1.2.2+1
      noty1.2.2[count1.2.2]←y1.2[i]
   }
}
y1.1.1←noty1.1.1[1:count1.1.1]
y1.1.2←noty1.1.2[1:count1.1.2]
y1.2.1←noty1.2.1[1:count1.2.1]
y1.2.2←noty1.2.2[1:count1.2.2]
count2.1.1←0
count2.1.2←0
noty2.1.1←c(rep(0,n2.1))
noty2.1.2←c(rep(0,n2.1))
for (i in 1:n2.1)
{
   if (z2.1[i]==1)
   {
      count2.1.1←count2.1.1+1
      noty2.1.1[count2.1.1]←y2.1[i]
   }
   else
   {
      count2.1.2←count2.1.2+1
      noty2.1.2[count2.1.2]←y2.1[i]
   }
}
count2.2.1←0
count2.2.2←0
noty2.2.1←c(rep(0,n2.2))
noty2.2.2←c(rep(0,n2.2))
for (i in 1:n2.2)
{
   if (z2.2[i]==1)
   {
      count2.2.1←count2.2.1+1
      noty2.2.1[count2.2.1]←y2.2[i]
   }
   else
```

```
   {
      count2.2.2←count2.2.2+1
      noty2.2.2[count2.2.2]←y2.2[i]
   }
}
y2.1.1←noty2.1.1[1:count2.1.1]
y2.1.2←noty2.1.2[1:count2.1.2]
y2.2.1←noty2.2.1[1:count2.2.1]
y2.2.2←noty2.2.2[1:count2.2.2]
minusloglike1←function(par)
{
   sum1.1.1←sum(log(pnorm(par[1]-par[3]-par[2]*y1.1.1)))
   sum1.1.2←sum(log(1-pnorm(par[1]-par[3]-par[2]*y1.1.2)))
   sum1.2.1←sum(log(pnorm(par[1]-par[2]*y1.2.1)))
   sum1.2.2←sum(log(1-pnorm(par[1]-par[2]*y1.2.2)))
   minusloglike1←-(sum1.1.1+sum1.1.2+sum1.2.1+sum1.2.2)
}
minusloglike2←function(par)
{
   sum2.1.1←sum(log(pnorm(par[1]-par[3]-par[2]*y2.1.1)))
   sum2.1.2←sum(log(1-pnorm(par[1]-par[3]-par[2]*y2.1.2)))
   sum2.2.1←sum(log(pnorm(par[1]-par[2]*y2.2.1)))
   sum2.2.2←sum(log(1-pnorm(par[1]-par[2]*y2.2.2)))
   minusloglike2←-(sum2.1.1+sum2.1.2+sum2.2.1+sum2.2.2)
}
mle1←nlmin(f=minusloglike1,x=init1,max.iter=100,max.fcal=100)
est1.vec←mle1$x
mle2←nlmin(f=minusloglike2,x=init2,max.iter=100,max.fcal=100)
est2.vec←mle2$x
delta1←(p1.hat-p2.hat)*log(p1.hat/p2.hat)
+(q1.hat-q2.hat)*log(q1.hat/q2.hat)
delta2←(p1.hat+p2.hat)*(mu1.1.hat-mu2.1.hat)^2/(2*sigmasq.hat)
+(q1.hat+q2.hat)*(mu1.2.hat-mu2.2.hat)^2/(2*sigmasq.hat)
delta3←(p1.hat+p2.hat)*(est1.vec[3]-est2.vec[3])^2/2
delta←delta1+delta2+delta3
test.stat←(ssize1*ssize2*delta)/(ssize1+ssize2)
critical←qchisq(1-level,df=7)
if (test.stat>critical)
{
   count←count+1
}
else
```

```
    {
        count←count
    }
}
empirical.power←count/rep
```